In press at *Psychological Review*. **DOI: 10.1037/rev0000126**

Language Learning as Language Use:

A Cross-linguistic Model of Child Language Development

Stewart M. McCauley

Department of Psychological Sciences, University of Liverpool

Morten H. Christiansen

Department of Psychology, Cornell University

Short title: Language Learning as Language Use

Total words in main text: 27,062

Number of figures: 16 (in main text)

Corresponding author: Stewart M. McCauley
Department of Psychological Sciences
Eleanor Rathbone Building
74 Bedford St S
Liverpool
L69 7ZA
E: Stewart.McCauley@liverpool.ac.uk
P: +44 (0)151 794 1111

In press at *Psychological Review*. **DOI: 10.1037/rev0000126**

Abstract

While usage-based approaches to language development enjoy considerable support from computational studies, there have been few attempts to answer a key computational challenge posed by usage-based theory: the successful modeling of language learning as language *use*. We present a usage-based computational model of language acquisition which learns in a purely incremental fashion, through on-line processing based on chunking, and which offers broad, cross-linguistic coverage while uniting key aspects of comprehension and production within a single framework. The model's design reflects memory constraints imposed by the real-time nature of language processing, and is inspired by psycholinguistic evidence for children's sensitivity to the distributional properties of multi-word sequences and for shallow language comprehension based on local information. It learns from corpora of child-directed speech, chunking incoming words together to incrementally build an item-based "shallow parse." When the model encounters an utterance made by the target child, it attempts to generate an identical utterance using the same chunks and statistics involved during comprehension. High performance is achieved on both comprehension- and production-related tasks: the model's shallow parsing is evaluated across 79 single-child corpora spanning English, French, and German, while its production performance is evaluated across over 200 single-child corpora representing 29 languages from the CHILDES database. The model also succeeds in capturing findings from children's production of complex sentence types. Together, our modeling results suggest that much of children's early linguistic behavior may be supported by item-based learning through on-line processing of simple distributional cues, consistent with the notion that acquisition can be understood as learning to process language.

Keywords: Language Acquisition; Corpora; Chunking; Shallow Parsing; Usage-Based Approach

Introduction

The ability to comprehend and produce an unbounded number of novel utterances has long been regarded as a hallmark of human language. How does a child acquire such productivity, given input that is both noisy and finite? For over half a century, generative linguists have argued that such open-endedness can only be explained by a system of abstract grammatical rules operating over word classes, scaffolded by innate, language-specific knowledge (e.g., Chomsky, 1957; Pinker, 1999). In recent years, however, an alternative theoretical perspective has emerged in the form of usage-based approaches (e.g., Croft, 2001; Goldberg, 2006; Tomasello, 2003), which hold that children's language development is initially item-based. Rather than being guided by system-wide abstract principles, productivity is taken to emerge gradually, beginning with concrete items in the child's input. This perspective is motivated in part by analyses of child-directed speech, showing that there is considerably more information available in the input than previously assumed (e.g., Redington, Chater, & Finch, 1998; Monaghan & Christiansen, 2008), as well as a wide range of observational and empirical work showing that children can use such information in an item-based manner. Such evidence includes cross-linguistic findings of item-specific patterns in early verb usage (e.g., Berman, 1982; MacWhinney, 1975; Gathercole, Sebastián, & Soto, 1999; Pizutto & Caselli, 1992; Rubino & Pine, 1998), as well as studies of children's production of novel verbs (e.g., Tomasello & Brooks, 1998; Akhtar, 1999), use of determiners (e.g., Mariscal, 2008; Pine & Lieven, 1997), case marking errors (e.g., Kirjavainen, Theakston, & Lieven, 2009), production of complex sentence types (e.g., Diessel & Tomasello, 2005), and question formation (e.g., Dabrowska, 2000).

In addition to this wealth of observational and empirical evidence, a number of computational modeling studies have provided a source of complementary support for usage-based approaches, using item-based learning to successfully capture specific developmental patterns (Freudenthal, Pine, & Gobet, 2006, 2007; Gobet, Freudenthal, & Pine, 2004; Jones, Gobet, & Pine, 2000), the acquisition of

item-based constructions and schemas (Chang, 2008; Solan, Horn, Ruppin, & Edelman, 2005), and semantic role learning (e.g., Alishahi & Stevenson, 2010), as well as tracing the emerging complexity of children's grammatical knowledge more generally (e.g., Bannard, Lieven, & Tomasello, 2009; Borensztajn, Zuidema, & Bod, 2009).

Despite the considerable success of item-based computational approaches to acquisition, there have been few computational accounts of the on-line processes driving children's attempts to comprehend and produce speech, or the ways in which these specific usage events incrementally contribute to the child's emerging linguistic abilities. This lack seems to stem, in part, from traditional ways of idealizing the task facing the language learner: from a computational standpoint, the issue of linguistic productivity tends to be approached primarily as a problem of grammar induction; to attain open-ended productivity, the learner must first identify a target grammar on the basis of exposure to a sample of sentences generated by that grammar (Gold, 1967). While computational approaches to acquisition have largely moved beyond Gold's formal learnability approach, incorporating a variety of different sources of linguistic information, the idealization of the task facing the learner as one of grammar induction has remained largely intact. As a consequence, computational work within the usage-based tradition has continued to focus on grammar induction (e.g., Borensztajn et al., 2009).

Usage-based theory suggests the possibility of sidestepping the grammar induction approach altogether, focusing instead on the ways in which linguistic knowledge is built up and reinforced through specific usage events (the child's attempts to comprehend and produce speech). This perspective has recently been bolstered by a number of complementary experimental results which suggest that the task facing learners is better characterized as one of “learning by doing” than as one of grammar induction (see also Chater & Christiansen, 2010, in press; Christiansen & Chater, 2016a). Evidence for the psychological reality of multiword linguistic units has served to blur the lines between grammar and lexicon, demonstrating the storage of “compositional” phrases as well as their use in

comprehension and production (e.g., Arnon & Snider, 2010; Bannard & Matthews, 2008; see also contributions in Christiansen & Arnon, 2017). Moreover, work on associative learning (e.g., Perruchet, Vinter, Pacteau, & Gallego, 2002) and statistical learning (e.g., Thompson & Newport, 2007) suggests that computationally simple mechanisms may be sufficient to identify the boundaries of such units in the speech stream.

A highly relevant—though previously unconnected—line of research has focused on the issue of syntactic processing depth, providing evidence that comprehension processes are often shallow and underspecified (e.g., Sanford & Sturt, 2002). Taken together with evidence for the primacy of local information during processing (e.g., Tabor, Galantucci, & Richardson, 2004), this suggests that children and adults form representations which are merely "good enough" for the communication task at hand (e.g., Ferreira & Patson, 2007). Evidence for multiword linguistic units, shallow processing, and the use of local information makes contact with other work emphasizing the importance of sequential as opposed to hierarchical linguistic structure (e.g., Culicover, 2013; Frank & Bod, 2011; Frank & Christiansen, in press; O'Grady, 2015; see Frank, Bod, & Christiansen, 2012, for a review).

Despite the importance of these complementary areas of research for strengthening item-based approaches, as well as their implications for re-characterizing the task facing language learners, they have remained largely unconnected. A recent theoretical proposal by Christiansen and Chater (2016b) unites these seemingly disparate strands of evidence. The proposal rests on the uncontroversial acknowledgement that language takes place in the "here and now." The consequences of this real-time constraint—which Christiansen and Chater refer to as the "Now-or-Never bottleneck"—are rarely considered, however. The fleeting nature of signal and memory have implications for how we approach human language: At a normal rate of speech, humans produce between 10 and 15 phonemes per second (Studdert-Kennedy, 1986). Nevertheless, the ability to process discrete sounds appears to be limited to about 10 items per second (Miller and Taylor, 1948), beyond which they are perceived to fuse into a

single buzzing sound. To make matters worse, the auditory trace is limited to about 100 ms (Remez, Ferro, Dubowski, Meer, Broder, & Davids, 2010). Moreover, memory for arbitrary sequences seems to be limited to about four items (Cowan, 2001; Warren, Obusek, Farmer, & Warren, 1969).

Thus, the signal—and human memory for it—are incredibly short-lived. On the surface, the Now-or-Never bottleneck would seem to render language learning and use impossible. A key strategy for overcoming these sensory and memory limitations lies in *chunking*: incoming items can be rapidly grouped and passed to successively higher levels of representation, with higher-level representations allowing input to be dealt with before it is overwritten by the onslaught of incoming information at a lower level. It is fairly intuitive and uncontroversial, for instance, that the raw acoustic signal is rapidly packaged into some sort of sound-based unit (e.g., phoneme- or syllable-like representations), which can in turn be chunked into word-like representations, and so on. The consequences of applying this general approach to sentence-level processing and grammatical development are, however, less obvious, as discussed by Christiansen and Chater (2016a, b).

Though gaining renewed emphasis under this perspective, chunking has been regarded as a key learning and memory mechanism in human cognition for over half a century (e.g., Feigenbaum & Simon, 1962; Miller, 1956; Simon, 1974). While verbal theories have been more common, computational models of chunking have been present in the literature for over four decades (e.g., Ellis, 1973; French, Addyman, & Mareschal, 2011; Jones, 2012; Perruchet & Vinter, 1998; Servan-Schreiber & Anderson, 1990; Simon & Gilmartin, 1973). Previous computational accounts of chunking have had a significant impact on approaches to language development, particularly with respect to the area of speech segmentation (cf. Frank, Goldwater, Griffiths, & Tenenbaum, 2010).

In what follows, we present a computational framework which extends the real-time use of chunking beyond word segmentation to aspects of sentence comprehension and production, uniting evidence for multiword linguistic units and shallow processing within a simple, developmentally

motivated model of acquisition that learns through on-line processing. We begin by discussing these lines of research as they pertain to our computational approach, before introducing the model and its inner workings. We then report results on the acquisition of English as well as the simulation of a key psycholinguistic experiment on children's sentence processing. Finally, we demonstrate that our approach extends beyond English to cover the acquisition of a broad array of typologically diverse languages.

The Psychological Reality of Multiword Linguistic Units

Our computational approach to acquisition begins with the idea that language learners form representations of differing granularities, with linguistic units ranging from the fine-grained level of morphemes and words to the more coarse-grained level of word sequences comprising one or more phrases. This perspective emerges straightforwardly from item-based approaches to acquisition; at the heart of usage-based theory lies the idea that linguistic productivity develops gradually through abstraction over multiword sequences (e.g., Abbot-Smith and Tomasello, 2006; Tomasello, 2003), requiring that storage of multiword units (chunks) occurs. In contrast, generative approaches have traditionally remained faithful to a words-and-rules perspective, in which learning and processing are supported by separate systems for lexicon and grammar (e.g., Pinker, 1999)[1].

While the assumption that children in some sense store multiword sequences has received support from naturalistic observation (e.g., Peters, 1983) and corpus analyses (e.g., Lieven, Behrens, Speares, & Tomasello, 2003), it is only recently that its validation has been made the target of experimental work. The finding of Bannard and Matthews (2008) that young children repeat phrases faster and more accurately when they form a frequent chunk may have provided the first direct

1More recent accounts, however, have allowed for storage of multiword sequences within a generative framework (Culicover & Jackendoff, 2005; Culicover, Jackendoff & Audring, 2017; Jackendoff, 2002).

evidence not only that multiword chunk storage takes place, but that this storage can actively facilitate processing. Controlling for substring frequency, they contrasted repetition of four-word phrases in which the fourth word was of either high or low frequency, given the preceding trigram. Two and 3-year-olds were more likely to repeat a phrase correctly when its fourth word combined with the preceding trigram to form a frequent chunk, while 3-year-olds were significantly faster to repeat the first three words. Further evidence comes from children's production of irregular plurals: Arnon and Clark (2011) found that the overregularization errors are significantly reduced when irregular plurals are produced in the context of lexically-specific frames (e.g., "*brush your* teeth").

The importance of such findings to usage-based approaches is underscored by previous computational modeling work demonstrating that the alignment and comparison (cf. Edelman, 2008) of multiword sequences can give rise to a considerable amount of linguistic productivity, through the abstraction of partially item-based grammatical constructions (Kolodny, Lotem, & Edelman, 2015; Solan et al., 2005).

While usage-based theory has focused primarily on the importance of stored sequences as exemplars in the abstraction of grammatical regularities, children's apparent use of multiword units during on-line processing (Arnon & Clark, 2011; Bannard & Matthews, 2008) highlights an active role for such units in comprehension and production, suggesting the possibility that multiword sequences retain their significance throughout development. Indeed, a number of findings indicate that the storage and active use of multiword units persists beyond early acquisition and into adulthood. Bannard and Ramscar (2007) found an effect of overall sequence frequency on reading times for units ranging from 4 to 7 words in length, while Reali and Christiansen (2007) showed chunk frequency effects in the processing of complex sentence types. Arnon and Snider (2010) found the same general pattern using a phrasal-decision task, whereby four-word expressions were classified as possible or impossible strings in English (in a vein similar to lexical-decision tasks). Importantly, Arnon and Snider's study explored

multiple frequency bins; reaction times decreased as a function of phrase frequency. Caldwell-Harris, Berant, and Edelman (2012) extended this finding to a broader frequency spectrum, showing a continuous effect of frequency and providing evidence against a frequency "threshold" beyond which a sequence is unitized. Additional evidence for adults' sensitivity to multiword sequence frequency has been gained from self-paced reading and sentence recall tasks (Tremblay, Derwing, Libben, & Westbury, 2011), eye-tracking data (Siyanova-Chanturia, Conklin, & van Hueven, 2011), and event-related brain potentials (Tremblay & Baayen, 2010). A similar pattern of results has been found in studies of adult production, demonstrating a decrease in naming latencies with increasing phrase frequency (Janssen & Barber, 2012) as well as reduced phonetic duration for frequent multiword sequences in elicited and spontaneous speech (Arnon & Cohen Priva, 2013)[2].

There are also direct parallels between the learning and processing of multiword units and individual words with respect to age-of-acquisition (AoA) effects. In a variety of tasks, adults exhibit processing advantages for words that are acquired earlier in childhood (for reviews, see Ghyselinck, Lewis, & Brysbaert, 2004; Johnston & Barry, 2006; Juhasz, 2005). Arnon, McCauley, and Christiansen (2017) show that multiword sequences, like individual words, display AoA effects when AoA is determined using either corpus-based metrics or subjective AoA ratings. They also show that the effect cannot be reduced to frequency, semantic plausibility, or lexical AoA. By underscoring a further parallel between words and multiword patterns, this study builds strong support the notion of stored multiword sequences as key building blocks for language learning and use.

Thus, the importance of multiword linguistic units extends beyond merely serving as exemplars for the formation of item-based schemas or the abstraction of grammatical regularities; multiword sequences play an active role in on-line processing, and this persists into adulthood. Accordingly, the

2 While the above studies have focused primarily on distributional properties, there are additional semantic and prosodic contributions to the ways in which language users represent and draw upon multiword units (e.g., Jolsvai, McCauley, & Christiansen, 2013).

on-line discovery and use of multiword sequences during comprehension and production forms one of the key features of the present computational approach.

The use of multiword linguistic units also leads us to explore the possibility that children's language development does not inevitably arrive at the use of fully articulated, hierarchical phrase structure, as assumed in many previous computational studies. In blurring the lines between lexicon and grammar, the active use of multiword sequences points to a potential role for relatively "flat" syntactic structures, suggesting that a more shallow form of processing may persist throughout development and into adulthood. This radically changes the problem facing the learner; instead of being forced to learn global hierarchical structures tied to a target grammar, local sequential structure moves to the fore. In what follows, we explore this idea more closely, reviewing evidence that shallow processing based on local information represents the norm rather than the exception in language use.

The Ubiquity of Shallow Processing in Language Use

Evidence for shallow processing of linguistic input has led some researchers to question the centrality of hierarchical phrase structure as well as the standard generativist assumption that syntactic and semantic processes are carried out completely and automatically. Yet for over half a century, hierarchical phrase structure has been viewed as a key theoretical foundation of most accounts of language acquisition and processing (e.g., Chomsky, 1957). Consequently, the idea that the meaning of a sentence need not stem from a fully articulated syntactic structure remains controversial. Nevertheless, shallow processing has been shown to be a widespread phenomenon through psycholinguistic research (for reviews, see Ferreira, Bailey, & Ferraro, 2002; Sanford & Sturt, 2002), and while the vast majority of this work has dealt with adult subjects, the theoretical implications extend from adult processing to the study of language acquisition. Here, we briefly discuss the evidence for shallow processing in adult language users before turning our attention to similar (though

much more limited) evidence from developmental studies, and finally outlining an account of shallow sentence processing which forms part of the motivation for the computational approach to acquisition put forth in this paper.

What is perhaps the most well-known thread of evidence for shallow processing comes from the failure of readers to notice semantically anomalous words and phrases in texts, indicating that processes of semantic integration have not been fully completed by readers who nevertheless form coherent semantic representations based on the sentences in question (e.g., Barton & Sanford, 1993; Erickson & Mattson, 1981). Other work has focused on text-change blindness (following work on change blindness in visual processing; e.g., Simons & Levin, 1998) to demonstrate the extent to which several factors modulate depth of processing, including focus (Sanford, 2002; Sturt, Sanford, Stewart, & Dawydiak, 2004) and computational load (Sanford, Sanford, Filik, & Molle, 2005). Perhaps more relevant is work demonstrating subjects' interpretation of nonsensical sentences as coherent (Fillenbaum, 1974; Wason & Reich, 1979) as well as the processing of semantically anomalous sentences in ways that directly contradict the interpretations that would be made according to a full syntactic parse (Ferreira, 2003), demonstrating the on-line use of background world knowledge and pragmatic expectations.

The above-mentioned evidence for shallow processing meshes naturally with work highlighting readers' tendencies to form “underspecified” representations of sentences, in which no commitment is made to any one of a number of possible analyses, clearly indicating that fully articulated syntactic processing has not taken place. Evidence for underspecification comes from work involving ambiguous relative clause attachment (Swets, Desmet, Clifton, & Ferreira, 2008), quantifier scope (Tunstall, 1998), metonymy (Frisson & Pickering, 1999), ambiguous nouns (Frazier & Rayner, 1990), and anaphoric reference (Koh, Sanford, Clifton, & Dawydiak, 2008). Like shallow processing more generally, underspecified representations are at odds with theories of processing that assume full

completion of syntactic and semantic analyses. Much of the evidence for underspecification makes good contact with Ferreira and Patson's (2007) Good Enough approach to sentence processing, in which it is argued that the goal of language comprehension is to establish representations which are merely "good enough" to suit the needs of a listener or reader in a given situation, as opposed to representing communicated information in full detail[3].

Taken together, the evidence suggests that shallow, underspecified processing, far from representing a degenerate case or mere exception to normal full syntactic and semantic processing, is ubiquitous. It is worthy of note that current evidence for shallow processing comes from work with written texts, a medium which allows subjects to process language without facing considerable challenges from 1) the highly noisy, variable nature of the speech signal and 2) the time constraints that come with not being able to control the speed at which input is encountered[4]. Thus, it is likely that much stronger evidence for shallow processing can be gained using speech stimuli (cf. Christiansen & Chater, 2016b).

In the above-mentioned cases, readers seem to rely on local linguistic information and global background knowledge rather than compositional meanings derived from fully articulated syntactic representations. Thus, support for shallow processing makes close contact with the claim that adults process sentences by using small chunks of local information to arrive at a semantic representation (e.g., Ferreira & Patson, 2007), which is reflected by local coherence effects (e.g., Tabor et al., 2004).

Evidence that adults process sentences in this manner makes it likely that children may rely on

3 As pointed out by Sanford and Sturt (2002), the contrast between traditional notions of full syntactic processing and shallow, underspecified processing is mirrored in the fields of computational linguistics and natural language processing (NLP) by differences between the output of shallow parsers, which identify a subset of interpretations for a sentence, and full syntactic parsers, which build a fully articulated syntactic analysis. Even in the context of NLP, shallow parsing sometimes offers computational advantages over full parsing (e.g., Li & Roth, 2001). Recently, it has also been shown that shallow parsing is sufficient for semantic role labeling in a morphologically rich language (Goluchowski & Przepiorkowski, 2012).

4 Note that there may also be strict time pressures during normal fluent reading, when readers take in about 200 words per minute (see Chater & Christiansen, 2016, for discussion).

similarly shallow, underspecified processing in which local information is key. While the issue of syntactic processing depth remains largely unexplored in children, initial evidence suggests that young learners rely upon shallow, underspecified processing to an even greater extent than adults (e.g., Gertner & Fisher, 2012). Corpus analyses of child speech similarly suggest that children's earliest complex sentences featuring sentential complements (e.g., *I think I saw one*) represent the simple concatenation of a formulaic expression (*I think*) with a sentence (*I saw one*) in a shallow rather than hierarchical fashion (Diessel & Tomasello, 2000).

Evidence for shallow processing based on local information makes close contact with Sanford and Garrod's (1981, 1998) Scenario Mapping and Focus theory of comprehension, in which background knowledge of situations and scenarios is used on-line to interpret linguistic input as it is encountered. During on-line interpretation, incoming linguistic input is mapped onto schemas of events, situations, or scenarios which have been established based on previous contexts or input – interpretation of the overall message is therefore heavily influenced by the background information which linguistic input is mapped onto. It may therefore be fruitful to test the view of language comprehension as the attempt to map chunks of language input onto specific parts of a scenario or event schema (which can, of course, be quite abstract and need not correspond to concrete objects and events in the real world); shallow processing may be sufficient for accomplishing this task. This, in turn, helps us reframe the problem facing the language learner: multiword unit learning (which allows rapid and efficient retrieval of chunks of local information during comprehension and production) naturally dovetails with a shallow processing approach, allowing language learners to comprehend much of the input without the need for full global syntactic parsing of the sort assumed in the vast majority of approaches to language learning.

This perspective fits nicely with several threads of psycholinguistic and computational work which are beginning to converge on the view that language users rely on sequential rather than

hierarchical structures (for a review, see Frank et al., 2012; Frank & Christiansen, in press). For instance, Ferreira and Patson (2007) found that interpretations can be constructed on the basis of small numbers of adjacent items, at the expense of more global syntactic structures and meanings, suggesting that global hierarchical structures were either impeded by local information or were altogether less important. Along the same lines, Christiansen and MacDonald (2009) found that simple recurrent networks (Elman, 1990), which simply learn to predict upcoming items in sentences in a linear rather than hierarchical fashion, provide a close fit to the abilities of human subjects to process recursive constructions involving center-embedding or cross-dependency. Consistent with this finding, Frank and Bod (2011) demonstrated that models which learn linear, non-hierarchical sequential information about word classes provide a stronger fit to actual human eye movement data during reading than models which learn hierarchical phrase structures.

In line with the view that sentence processing relies heavily on sequential structures computed over chunks of local information, our computational approach is centered on simple mechanisms for the on-line discovery, storage, and sequencing of words and chunks through sensitivity to the local rather than global information contained in utterances. Before detailing our computational approach in greater depth, we briefly discuss the potential sources of information children might use to discover useful chunks of local information, and the relationships between them, during their attempts to comprehend and produce utterances.

Discovering Useful Multiword Sequences

Our computational account of children's on-line processing seeks to capture some of the mechanisms by which multiword units are learned and employed in language comprehension and production. For the sake of simplicity, we distinguish between two types of multiword units: 1) *unanalyzed chunks*, and 2) *combined chunks* (see also Arnon & Christiansen, 2017; McCauley, Monaghan & Christiansen,

2015). Unanalyzed chunks are those that are stored and accessed holistically before segmentation of their parts has taken place, whereas combined chunks can be (and sometimes are) broken down into their component words. Chunks falling into the first category are most relevant to the study of very early language development (for a recent incremental, on-line computational model of word segmentation which captures processes whereby unanalyzed chunks can be discovered and gradually broken down into smaller units, see Monaghan & Christiansen, 2010). As an example, the chunk *look at this* may be treated as a holistic, unanalyzed unit by very young children (for a review of the literature on children's use of such "frozen" sequences, see Wray, 2005), while the same chunk would fall into the second category (as a *combined chunk*) for older children who are capable of breaking the chunk down into its component parts.

Beyond a certain point, chunks will rarely be treated as holistic units: Consider evidence that idioms, which even generative grammar-oriented approaches recognize as stored (Jackendoff, 1995; Pinker, 1999), prime, and are primed by, their component words (Sprenger, Levelt, & Kempen, 2006) as well as lead to syntactic priming (Konopka & Bock, 2009). Given that idioms would appear to form stored multiword units (their meanings are idiosyncratic and cannot be determined on the basis of component parts), we must allow, then, that a multiword unit can be accessed and used as a meaningful entity in its own right, even when activation of its individual parts occurs.

One well-studied source of information that infants might use to arrive at some of their earliest, unanalyzed multiword chunks lies in the acoustic correlates of clause and phrase boundaries. Pre-linguistic infants can use prosodic information to segment the speech stream into multiword units, and this ability has been shown to facilitate certain types of processing. Early work in this vein established that infants are sensitive to the prosodic correlates of clause boundaries (Hirsh-Pasek, Kemler Nelson, Jusczyk, Cassidy, Druss, & Kennedy, 1987). Further work demonstrated that infants are better able to recall phonetic information when it is packaged in a prosodically well-formed unit, and that infants can

use the acoustic correlates of clause boundaries to form representations which are available for use in later segmentation of continuous speech (Mandel, Jusczyk, & Kemler-Nelson, 1994). More relevant to the present study is work on phrase-level units. Though several studies suggest that phrases are not as reliably marked in the speech stream as are clauses (Beckman & Edwards, 1990; Fisher & Tokura, 1996), infants' sensitivity to these markers has been demonstrated (Jusczyk, Hirsh-Pasek, Kemler Nelson, Kennedy, Woodward, & Piwoz, 1992). Moreover, it has been shown that infants can use these markers to segment larger prosodic units corresponding to clauses into smaller, phrase-level units (Soderstrom, Seidl, Kemler-Nelson, & Jusczyk, 2003). Results from the Soderstrom et al. (2003) study went beyond mere on-line recognition of prosodic ill-formedness, suggesting that infants formed representations based on the prosodic information in familiarization sequences, and used them to segment prosodically well-formed items into phrase-level units at test.

The incorporation of such prosodic information, however, represents a challenge for computational models of acquisition, given the limited availability of prosodic information in currently available corpora of child-directed speech. Fortunately, distributional information is also highly relevant to early chunk discovery. Some of children's earliest unanalyzed multiword chunks may stem from “errors” in word segmentation (as suggested by Bannard & Matthews, 2008). For instance, using mutual information between syllables to find word boundaries in an unsegmented corpus, Swingley (2005) found that 91% of bisyllabic false alarms were frequent word pairs, such as *come on,* while 68% of trisyllabic false alarms were frequently occurring multiword phrases. More recent models of word segmentation (e.g., Goldwater, Griffiths, & Johnson, 2009; Monaghan & Christiansen, 2010) have yielded similar results. Given that such models exploit some of the same distributional cues that infants have been shown to be sensitive to in experimental studies of artificial word segmentation, it would not be surprising if infants similarly undersegmented the speech stream to arrive at unanalyzed multiword chunks. Far from hindering the child's language development, such “mistakes” may actually impart an

advantage, as predicted by usage-based theories (e.g., Arnon, 2009; Arnon & Christiansen, 2017).

But what of chunks acquired after segmentation of the component words has taken place? Presumably, unanalyzed chunks, arrived at through under-segmentation and/or the use of prosodic information, are rare once the child reaches a certain level of experience. The usefulness of multiword chunks should be no less real for an experienced language user (and indeed, as shown in the studies reviewed above, older children and adults actively use multiword units). Thus, we should allow for the possibility that statistical information linking words can be used to arrive at multiword chunks by older children and adults.

How might learners chunk co-occurring words together as a unit after segmentation of the component parts has already taken place? The use of raw frequency of co-occurrence would lead to placing too much emphasis on co-occurring words that frequently occur adjacent to one another by mere virtue of being highly frequent words. Similarly, precise tracking of the frequencies of all encountered sequences would lead to a combinatorial explosion (cf. Baayen, Hendrix, & Ramscar, 2013).

Thus, while phrase-frequency effects are continuous rather than threshold-based (e.g., Caldwell-Harris et al., 2012), meaningful chunks cannot be identified on the basis of raw frequency alone[5], in much the same way as a word segmentation model based solely on raw frequency of co-occurrence would be largely ineffective. Consistent with the Now-or-Never perspective (Christiansen and Chater, 2016b), which forms part of the theoretical motivation for the present study, we explore the notion that many of the same cues and mechanisms involved in word segmentation may be involved in chunking at the level of multiword units. In what follows, we discuss previous computational accounts of chunking

[5] Indeed, it has been suggested that part of the problem experienced by second-language learners may be due to a suboptimal chunking strategy based on raw frequency (Ellis, Simpson-Vlach, & Maynard, 2008)—something that has been corroborated by simulations of second-language learning using the CBL model presented below (McCauley & Christiansen, 2017).

which, in the domain of language, have been primarily concerned with word segmentation. We then describe our own model, which extends chunk-based learning and processing to the sentence level.

Previous Computational Models of Chunking

Chunking has been regarded as a key learning and memory mechanism for over half a century (e.g., Feigenbaum & Simon, 1962; Miller, 1956), with many of the earliest computational implementations being concerned with expertise (e.g., Ellis, 1973; Simon & Gilmartin, 1973) or language-related phenomena such as spelling (Simon and Simon, 1973), alphabet recitation (Klahr, Chase, & Lovelace, 1983), and statistical learning (Christiansen, in press). In recent decades, a number of chunking models related to implicit learning have emerged, and have been applied to word segmentation, particularly in the context of modeling data from artificial language learning experiments.

An early instance of one such model is the Competitive Chunking (CC) model of Servan-Schreiber and Anderson (1990), which views learning as the buildup of progressively larger chunks which are structured in a hierarchical network. Servan-Schreiber and Anderson argued that the implicit learning of artificial grammars (e.g., Reber, 1967) is primarily driven by chunking, and model the discrimination of grammatical vs. ungrammatical strings according to the number of stored chunks necessary to describe a sequence. The CC model operates according to activation of hierarchical chunks which match a current stimulus. Activated chunks which overlap with one another then "compete" to shape perception of the stimulus. Chunk creation and retrieval are determined by chunk strength, which is tied to free parameters involving decay and competition. In a Reber (1967) task analogue, CC was able to capture 87% of the variance in subject discrimination of grammatical vs. ungrammatical strings.

Perhaps the most influential chunking model devoted to implicit learning and word segmentation is PARSER (Perruchet & Vinter, 1998), which was directly inspired by the CC model of

Servan-Schreiber and Anderson (1990). Unlike the CC model, however, PARSER does not build up a hierarchical network of chunks, being concerned primarily with identifying structurally relevant units, such as words (Perruchet et al., 2002). Like the CC model, PARSER operates through competition, or "interference," between overlapping chunks, and utilizes free parameters for setting decay and managing activation rates (chunk strength). PARSER has been used to successfully model some of the experimental data involving human word segmentation performance in artificial language learning contexts (e.g., Saffran, Newport, & Aslin, 1996; cf. Perruchet & Vinter, 1998), and has also been used to discover the syntactically relevant units in an artificial language generated by a finite-state grammar (Perruchet et al., 2002).

A more recent approach to chunking is the MDLChunker (Robinet, Lemaire, & Gordon, 2011), which operates according to the information-theoretic principle of minimum description length (MDL), following the notion that human cognition favors simpler representations as a general principle (Chater & Vitányi, 2003). Like CC, MDLChunker involves hierarchies of chunks. Unlike CC, or PARSER, MDLChunker does not have free parameters. MDLChunker captures human chunking in a novel task involving meaningless visual symbols, as well as providing similar results to PARSER and CC on a well-known experiment by Miller (1958).

Another contemporary model of perhaps greater relevance is the TRACX model (French et al., 2011). A connectionist auto-associator, TRACX operates according to recognition-based processing rather than prediction, as in the case of prediction-based networks like simple recurrent networks (SRNs; Elman, 1990). TRACX provides a better fit to human data than either PARSER or SRNs, across a range of sequence segmentation studies.

French et al. (2011) nicely exemplify the tensions in the implicit learning literature between recognition-based processing, through chunking, and statistically-based processing utilizing transitional probabilities (TPs). Despite the absence of predictive processing and lack of conditional probability

calculation, TRACX is sensitive to TPs in both directions, as is the case with PARSER.

A number of studies have argued that recognition-based chunking provides a better fit to human performance than do TPs: PARSER offers a better fit to human data than learning based solely on TPs in a context in which the statistics of two consecutively learned artificial languages are pitted against one another (Perruchet, Poulin-Charronnat, Tillmann, & Peereman, 2014). Moreover, PARSER provides a better fit to adult learning of a semi-artificial language than SRNs (Hamrick, 2014). Poulin-Charronnat, Perruchet, and Tillmann (2016) developed a design which allowed them to dissociate the influence of familiarity and transitional probabilities using a pre-exposure phase in a standard artificial word segmentation task, in which recognition of familiar units appeared to override sensitivity to statistical cues, though findings were only partially captured by PARSER. These findings are compatible with a recent recognition-based model of word segmentation (Monaghan & Christiansen, 2010) which performs reasonably well on child-directed speech input in comparison to more computationally complex models.

Of the above computational approaches to chunking, PARSER has been the most widely explored in the context of human artificial-language data on chunking and segmentation performance. PARSER also best satisfies the memory constraints imposed by the Now-or-Never bottleneck, which forms part of the theoretical motivation for the present study. MDLChunker, for instance, has no memory limitations and is ill-suited to capturing on-line processing. Therefore, we chose PARSER as a baseline comparison to our own model, alongside a standard, prediction-based model utilizing transitional probabilities over *n*-grams.

Integrating Recognition-based and Statistically-based Processing

While chunking models have been argued to offer a better fit than TPs to human performance in studies of artificial word segmentation, which involve brief periods of exposure, fewer studies have examined

learning over longer periods of time, or the learning of higher-level chunks, such as would be useful in learning grammatical regularities.

Transitional probabilities have been found to be useful in segmenting out multiword phrases: Thompson and Newport (2007) found that peaks in forward transitional probabilities (FTPs) between form classes in an artificial language can be used by adult subjects to group artificial words together into multiword units, whereas dips in FTPs can be used to identify chunk boundaries. However, there are a number of cases in which a sole reliance on forward transitional probabilities in natural language might prevent the segmentation of useful multiword chunks. For example, if learners were to compute statistics over individual words rather than form classes, the FTP between the words in an English phrase such as "*the dog*" will always be extremely low, given the sheer number of nouns that may follow a determiner. Other sources of information, however, such as mutual information or backwards transitional probabilities (BTPs) provide a way around this issue: given the word "dog," the probability that the determiner "the" immediately precedes it is quite high, considering the small number of determiners one might choose from. Thus, it makes sense that child learners might also make use of such sources of information to discover useful multiword units.

Along these lines, Saffran (2001, 2002) has shown that dependencies in the form of BTPs between words in an artificial phrase-structure grammar not only facilitate learning, but aid in isolating specific phrases. That infants and adults are sensitive to BTPs has been established in the chunking (French et al., 2011; Perruchet & Desaulty, 2008) and statistical learning (Pelucchi, Hay, & Saffran, 2009) literatures. Thus, the view of BTPs as a potential cue to useful multiword phrases holds promise. That English speakers may be more sensitive to BTPs than FTPs between words during production is suggested by auditory corpus analyses showing that functors and content words are shorter when predictable given the following word, while the same effect in the forward direction appears only for the most frequent functors, and is absent for content words (Bell, Brenier, Gregory, Girand, & Jurafsky,

2009). Though much previous work examining adults' production of multiword units has been concerned with raw frequency counts rather than conditional probabilities, the Bell et al. (2009) findings might also be taken as initial support for idea that speaker sensitivity to BTPs may help drive the representation and use of multi-word chunks.

The simple architecture of the model used in the present study—while inspired by the successes of recognition-based chunking in accounting for experimental data—also seeks to incorporate statistical learning, in line with the aforementioned evidence for the use of TPs in phrase-level chunking. Rather than rely on prediction-based processing, statistical information tied to BTPs is used as a cue for identifying chunks which are then stored as concrete units and used to support further processing through recognition. Forward prediction in the model is chunk-based rather than statistical. A purely recognition-based model which does not directly utilize statistical cues, PARSER (Perruchet & Vinter, 1998), serves as a baseline for comparison to our own model, alongside a purely prediction-based model utilizing transitional probabilities over *n*-grams.

The Chunk-Based Learner Model

In what follows, we present the Chunk-Based Learner (CBL) model of language learning. Following Christiansen and Chater (2016b), one of the primary aims of the CBL model is to provide a computational test of the idea that the discovery and on-line use of multiword units forms part of the backbone for children's early comprehension and production. To this end, the model gradually builds up an inventory of chunks consisting of one or more words—a “chunkatory”—which is used to simulate aspects of both language comprehension and production. The model explicitly captures the shallow processing perspective outlined above—in which chunks of local information are used to process sentences—by learning to group words together into local chunks that are appropriate for

arriving at an utterance's meaning (a key aspect of comprehension), while simultaneously learning to produce utterances in an incremental fashion using the same chunks of local information (a key aspect of production).

CBL was designed with several key psychological and computational features in mind:

1. **On-line processing**: During comprehension, input is processed word-by-word as it is encountered, reflecting the incremental nature of human sentence processing (e.g., Altmann & Steedman, 1988; Tanenhaus, Carlson, & Trueswell, 1989; Tyler & Marslen-Wilson, 1977); during production, utterances are constructed incrementally according to a chunk-to-chunk process rather than one of whole-sentence optimization (e.g., whereby many candidate sentences are compared simultaneously and the one with the highest score selected).This approach is consistent with memory constraints deriving from the real-time nature of language processing (Christiansen & Chater, 2016b).
2. **Incremental learning**: At any given point in time, the model can only rely on what it has learned from the input encountered thus far (i.e., unlike the vast majority of computational approaches to acquisition, the current model does not rely on batch learning of any sort).
3. **Simple statistics**: For reasons detailed above, learning is based on the computation of BTPs, which 8-month-old infants (Pelucchi, Hay, & Saffran, 2009) and adults (Perruchet & Desaulty, 2008) can track.
4. **Local information**: Learning is tied to local rather than global information; instead of storing entire utterances as sequences, the model learns about transitions between adjacent words and chunks.
5. **Item-based**: The model learns from concrete words, without recourse to abstract information such as that of syntactic categories (as is also the case with a number of other usage-based

models; e.g., Freudenthal et al., 2006; Jones et al., 2000; Kolodny et al., 2015; Solan et al., 2005). This stands in stark contrast to most computational approaches emerging from the tradition of generative linguistics; rule-based processing in the "words-and-rules" framework operates over word classes rather than words themselves.

6. **Psychologically motivated knowledge representation**: In accordance with evidence for the role of multiword linguistic units in comprehension and production (reviewed above) as well as for the interconnectedness of comprehension and production processes more generally (Chater, McCauley & Christiansen, 2016; Pickering & Garrod, 2007, 2013), aspects of both comprehension and production are performed by using the same inventory of single- and multiword linguistic units.
7. **Naturalistic input:** The model learns from corpora of child-directed speech taken from the CHILDES database (MacWhinney, 2000). As word segmentation itself lies outside the scope of the current model, the use of such pre-segmented corpora (which consist of words rather than phonemic transcriptions) enables us to expose the model to a far more diverse array of corpora than would be possible otherwise.
8. **Broad, cross-linguistic coverage**: The model is designed in such a way that it can be evaluated on corpora of child-directed speech in any language (following Chang, Lieven, & Tomasello, 2008). We therefore evaluate it using a typologically diverse set of 29 languages.

We begin by providing an initial glance at the general architecture of the model, before describing its inner workings in full detail. We then present results from simulations of the acquisition of English. Finally, we show that the model successfully extends to the acquisition of a wide array of typologically diverse languages.

Inner Workings of the Model

There is growing behavioral and neuroimaging evidence for the involvement of the same mechanisms in adult comprehension and production (for reviews, see Pickering & Garrod, 2007, 2013), which prompts us to extend the idea of a unified framework for comprehension and production to language development (Chater et al., 2016; McCauley & Christiansen, 2013). In light of this, we designed CBL to capture the idea that comprehension and production can be viewed as two sides of the same process, but with different task demands. Comprehension is approximated by the segmentation of incoming speech into chunks relevant for determining the meaning of an utterance. These units are then stored in an inventory that makes no distinction between single- and multi-word chunks. Production is approximated by the model's ability to reproduce actual child utterances through retrieval and sequencing of units discovered during comprehension. Crucially, the very same distributional information underlying the model's comprehension-related processing form the basis for production, while production itself is taken to feed back into comprehension.

The model's inventory of single- and multi-word linguistic units—its chunk inventory, or "chunkatory"—is its core feature. Comprehension-related processing is used to build up the chunkatory while production-related processing both draws upon it and reinforces it. Through the chunkatory, CBL is able to approximate elements of comprehension and production within a unified framework.

The model begins by learning—in an incremental, on-line fashion—to segment incoming input into groups of related words (similar to phrasal units). These chunks are then stored in the chunkatory unless they have been encountered before, in which case the frequency of the relevant chunk is incremented by 1. In each simulation, the input consists of a corpus of speech directed to a single child (taken from the CHILDES database; MacWhinney, 2000). When the model encounters a multiword utterance produced by the target child, the production side of the model comes into play: the model's task is to produce an utterance which is identical to that produced by the child, using only statistics and

chunks learned and used during comprehension-related processing up to that point in the simulation. Thus, we aimed to construct a fully incremental, on-line model of child language development that uses the same chunks and distributional statistics to perform aspects of both comprehension and production.

To summarize this initial snapshot of the model's inner workings: CBL approximates aspects of both *comprehension,* by learning an inventory of chunks and using them to segment child-directed speech into related groups of words (such as would be appropriate for arriving at an utterance's meaning via shallow processing), and *production*, by reproducing actual child utterances as they are encountered in a corpus, using the same chunks and statistics learned and used during comprehension. We hypothesized that both problems could, to a large extent, be solved by recognition-based processing tied to chunks which are discovered through sensitivity to *transitional probabilities* between linguistic units.

Transitional probability: The simple statistic at the heart of CBL: As reviewed above, TPs have been proposed as a cue to phrase structure in the statistical learning literature; peaks in TP can be used to group words together, whereas dips in TP can be used to find phrase boundaries (e.g., Thompson & Newport, 2007). The view put forth in such studies is that TP is useful for discovering phrase structure when computed over form classes rather than actual words. We hypothesized, instead, that distributional information tied to individual words provides richer source of information than has been assumed in such work. Because we adopted this purely item-based approach, and because of evidence for greater reliance on BTPs when chunking words together in English (Bell et al., 2009), we decided to focus initially on BTPs.

The computation of TPs in the backward direction also has an unexpected advantage in the context of incremental, on-line calculation, in that the properties of the most recently encountered word attain the greatest importance (e.g., the BTP linking the sequence XY can be arrived at by normalizing

P(X, Y) by P(Y) rather than involving X in the denominator when computing FTP). For the above-mentioned reasons, the current computational approach focuses initially on backward rather than forward TPs as a cue to multiword units—chunks of local information for use in processing—while comparing the types of TP in a systematic way in Appendix C.

In what immediately follows, we provide an in-depth description of the inner workings of the model, showing how simple TPs support recognition-based chunk learning through comprehension- and production-related processes.

Comprehension

Though comprehension and production in the model represent two sides of the same coin, we describe them separately for the sake of simplicity. During comprehension, the model discovers its first chunks through simple sequential statistics computed over words. Processing utterances on a word-by-word basis, the model learns frequency information for words and word pairs, which is used on-line to track BTPs between words and maintain a running average BTP across previously encountered word pairs. When the model calculates a BTP between two words that is greater than expected, based on the running average BTP, it groups the word pair together such that it may form part of a chunk. When the calculated BTP falls below the running average, a "boundary" is placed and the chunk thereby created—consisting of one or more immediately preceding words—is added to the chunkatory. Then, the model moves on to process the next word in the utterance. The use of the running average BTP as a threshold allows the avoidance of a free parameter. This process is illustrated using a simple utterance in Figure 1.

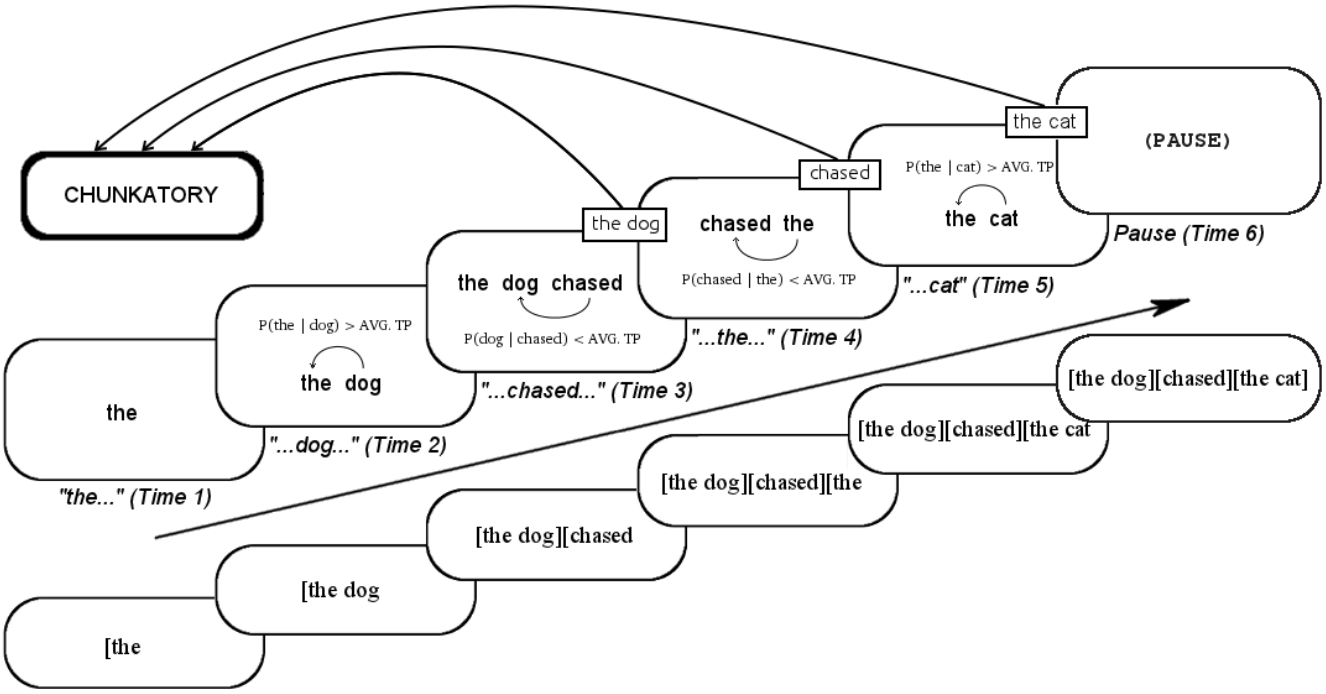


Fig. 1: Incremental, on-line processing of the simple utterance *"the dog chased the cat"*. Material above the diagonal arrow depicts the simple computations driving the model's on-line processing, while material below the arrow represents the resulting shallow parse (the model's interpretation of the sentence) as it unfolds over time. At Time 2, the model calculates the BTP between *the* and *dog*, which exceeds the average TP threshold (indicated by the backward arrow's position above the words), resulting in the two words being grouped together. Since the next word has not yet been encountered, the two words are not yet stored in the chunkatory as a chunk. At Time 3, the BTP between *dog* and *chased* falls below the running average (indicated by the backward arrow's position below the words), so *chased* is not grouped together with the preceding material and *the dog* is then stored in the chunkatory. At Time 4, the BTP between *chased* and *the* falls below the running average, so the two words are not grouped together and *chased* is added to the chunkatory as a single-word chunk. At Time 5, the BTP between *the* and *cat* rises above the average threshold and because a pause follows the sequence, *the cat* is chunked together and stored in the chunkatory.

All newly-added chunks are initialized with a frequency count of 1. The frequency count of a chunk is incremented by 1 each time it is encountered subsequently. Single-word utterances are

automatically treated as single chunks and stored (or their counts incremented) accordingly, though only multi-word utterances are scored when evaluating model performance (scoring procedures are described below).

Once the model has acquired its first chunk, it begins using its chunkatory in a recognition-based fashion to assist in processing the incoming input on the same incremental, word-to-word basis as before. The model continues learning the same low-level distributional information and calculating BTPs, but also uses the chunkatory to make on-line predictions as to which words should form a chunk, based on previously learned chunks. Crucially, these predictions are recognition-based rather than statistically-based. When a word pair is encountered, it is searched for in the chunkatory; if it has occurred more than once, either as a complete chunk or as part of a larger chunk, the words are automatically grouped together and the model moves on to the next word without placing a boundary. If the word pair has not occurred more than once in the chunks found in the chunkatory at that time step, the BTP is compared to the running average, with the same consequences as described above. Thus, there are no *a priori* limits on the number or size of chunks that can be learned.

As an example of how this can be understood as prediction, consider the following scenario in which the model encounters the phrase *the blue ball* for the first time and its chunkatory includes *the blue car* and *blue ball* (with frequency counts greater than 1). When processing *the* and *blue,* the model will not place a boundary between the two words because the word pair is already strongly represented in the chunkatory (as in *the blue car*). The model therefore predicts that this word-pair will form part of a chunk, even though the rest of the chunk has not yet been encountered. Next, when processing *blue* and *ball*, the model reacts similarly, as this word pair is also represented in the chunkatory. The model thereby combines its knowledge of two chunks to discover a new, third chunk, *the blue ball*, which is added to the chunkatory. As a consequence, the sequence *the blue* becomes even more strongly represented in the chunkatory, as there are now two chunks in which it appears.

Thus, once a chunk enters the chunkatory, it remains in the chunkatory. However, if the BTP linking the items in a chunk drops below the running average threshold before the model has re-encountered and "recognized" the chunk (i.e., before the chunk has attained a frequency count of 2 in the model's chunkatory), it will neither be incremented nor treated as a chunk. If the model has "recognized" the chunk in a previous encounter, it will continue to be considered as a unit (and its count incremented in the chunkatory), even if one of the BTPs linking the internal words should drop below the running average threshold. Importantly, there is no decay parameter in CBL.

Psychological validity of the model's multiword units: A recent study by Grimm, Cassani, Gillis, and Daelemans (2017) demonstrated that the CBL model extracts chunks with a uniquely facilitatory effect on child age-of-first-production, as well as adult reaction times in a word recognition task. That is, the greater the number of CBL-discovered chunks a word appears in, the earlier it is produced by children and the faster it is recognized by adults, even after controlling for the relevant variables (e.g., word frequency). This finding is further bolstered by comparing the CBL model to baseline models.

The CBL model's chunking mechanism therefore enjoys independent support from psychological findings related to children and adults. Moreover, as discussed below, the model has been used to successfully capture developmental psycholinguistic findings spanning a range of phenomena.

Production

While the model makes its way through a corpus incrementally, segmenting and storing chunks during comprehension, it encounters utterances produced by the target child, at which point the production side of the model comes into play. The model's ability to generate the child's utterance, based on chunks learned from previous input, is then evaluated using a sentence production task inspired by the

bag-of-words incremental generation task used by Chang et al. (2008), which offers a method for automatically evaluating syntactic learners on corpora in any language.

We loosely approximate the overall message that the child wants to convey by treating the utterance as an unordered set of words: a "bag-of-words," corresponding to (again, very roughly) the set of concepts contributing to the semantics of the utterance to be produced. The task for the model, then, is to sequence these words in the correct order, as originally produced by the target child. Following evidence for the use of multiword sequences in child production, as well as usage-based approaches more generally, the model utilizes its chunkatory to generate the child's utterances. In order to model retrieval of stored chunks during production, the bag-of-words is filled by comparing parts of the child's utterance against the chunkatory. For instance, consider a scenario in which the model is to produce the child utterance *the dog chased a cat* and the largest chunk in the chunkatory consists of 3 words. To begin, the first 3 words are searched for storage as a single chunk. As this is not found in the chunkatory, *the dog* is searched for. This search succeeds, so the words are removed from the utterance and placed in the bag as a single chunk. Next, *chased a cat* is searched for, unsuccessfully, followed by *chased a*, also without success. The word *chased* is placed in the bag as a single chunk. Then, *a cat* is searched for, and so on. Crucially, this procedure is not meant to correspond to a psychological process as such but simply used as a simulation shortcut to find chunks that the child already knows (i.e., that were in the chunkatory as a result of learning during comprehension) and thus would be likely to use as such (e.g., *the dog*). Once in the bag, the order of chunks is randomized.

During the second phase of production, the model attempts to reproduce the child's utterance using the unordered chunks in the bag-of-words. We model this as a gradual chunk-by-chunk process rather than one of whole-sentence optimization (e.g., calculating the probability of the entire utterance, etc.), in order to reflect the incremental nature of sentence processing (e.g., Altmann & Steedman,

1988; Christiansen & Chater, 2016b; Tanenhaus et al., 1989; Tyler & Marslen-Wilson, 1977). Thus, the model begins by removing from the bag-of-words the chunk with the highest BTP given the start-of-utterance marker (a hash tag representing the pause preceding the utterance in the corpus), and producing it as the start of its new utterance. The chunk is removed from the bag before the model selects and produces its next chunk, the one with the highest BTP given the previously produced chunk. In this manner, the model uses chunk-to-chunk BTPs to incrementally produce the utterance, adding chunks one-by-one until the bag is empty. The model's production of the child utterance *I'm gonna stop the train with my whistle* is depicted in Figure 2. In rare cases where two or more chunks in the bag-of-words are tied for the highest BTP, one of them is chosen at random.

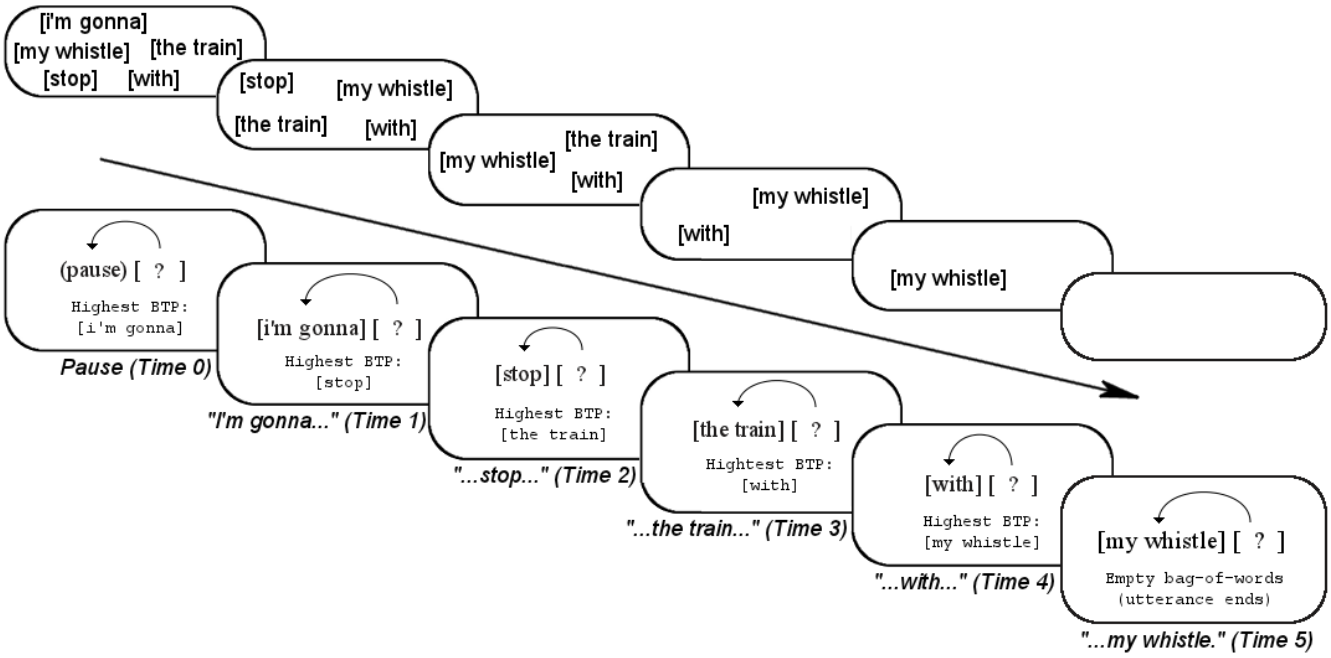


Fig. 2: Incremental, on-line production of the child utterance *"I'm gonna stop the train with my whistle."* Material above the diagonal arrow depicts the contents of the bag-of-words at each time step. Material below the arrow represents the simple computations whereby the model selects the next item to be produced at each time step. At

Time 0, the model selects its first chunk from the bag according to the highest BTP, given the pause preceding the utterance (which can be understood as a start-of-utterance marker); out of the chunks in the bag, [*i'm gonna*] has the highest BTP in this instance, so it is removed from the bag and produced at the next time step. At Time 1, the model calculates the BTP between [*i'm gonna*] and the remaining chunks in the bag; [*stop*] has the highest BTP and is therefore removed and produced at the next time step. This process continues, with the item possessing the highest BTP (given the previous item) being selected until the bag-of-words is empty, at which point the utterance ends.

Because comprehension and production are seen as two sides of the same process, a child's own productions are taken to reinforce statistics previously learned during comprehension. For this reason, immediately following the model's attempt to produce a given child utterance, the same utterance is used to reinforce the model's low-level sequential statistics as well as its chunkatory, through the performance of (incremental and on-line) comprehension on the utterance, in an identical manner to any other utterance of child-directed speech in the corpus. The child is taken to "hear" its own productions in a manner consistent with the position that no strong distinction can be drawn between the mechanisms and statistics underlying comprehension and production (as argued in Chater et al., 2016).[6] Thus, the CBL model features some similarities to the "Traceback Method" of Lieven et al. (2003), while also providing the kind of "rigorous computational evaluation" of the general approach called for by Kol, Nir, and Wintner (2014).

Validity of the bag-of-words production task: We recognize that there is more to language production than what is captured by the bag-of-word task, including the important contributions of semantics, pragmatics and real-world knowledge. Nonetheless, we suggest that the task does capture some important aspects of how distributional information may influence the sequencing of words

[6] While we found that the inclusion of this feedback mechanism did not lead to a statistically significant change in the model's shallow parsing performance, we feel that it is nevertheless a valuable inclusion on theoretical grounds. As discussed further in the General Discussion, we expect that a larger impact may be found in future versions of the model which incorporate semantics.

during sentence production. To ensure that our production task in this way approximates linguistic sequencing skills, we tested adult native speakers on a behavioral version of the task. Using the largest available child corpus of English (Maslen, Theakston, Lieven, & Tomasello, 2004), we extracted, at random, 50 grammatical child utterances and 20 child utterances which had been previously marked as ungrammatical, for a total of 70 test utterances. Twenty Cornell undergraduates (mean age 20.1 [SE 0.8], all native speakers of English) then received, for each utterance, an un-ordered set of chunks corresponding to the very same chunks the model used when attempting to produce the given utterance during a full simulation over the same corpus. The subjects' task, for each utterance, was to sequence the chunks to form a sentence. The mean accuracy rate was 95.6% across all subjects for the grammatical utterances, and 64% for the ungrammatical utterances (note that only a perfect match to the child's utterance was scored as accurate, just as with the model version of the task). Thus, we conclude that the bag-of-words task itself does provide a meaningful and valid test of linguistic skills which reflect knowledge of grammatical chunk sequences (had subjects performed at chance, this conclusion would be unwarranted).

Contrasting Recognition-based and Statistical Approaches to Chunking: Baseline Models

We compare CBL directly to two models: PARSER (Perruchet & Vinter, 1998) as well as a modified *n*-gram model. PARSER was chosen for comparison because it has been the most widely explored model in the context of human data on chunking and segmentation performance, while also best satisfying the memory constraints imposed by the Now-or-Never bottleneck (Christiansen & Chater, 2016b). As such, it provided an ideal instantiation of purely recognition-based processing for comparison to the alternative approach taken by the CBL model. As an additional baseline utilizing purely prediction-based processing, we implemented a variation on the standard *n*-gram model, focusing on trigrams for reasons explained below. The contrasting unit types and processing styles of

the model and its baselines are simplified in Table 1.

Table 1
Contrasting Unit Type and Processing Style

Model	Stored, Variable-sized Chunks?	Recognition-based?	Prediction-based?
CBL	Yes	Yes	Yes
PARSER	Yes	Yes	No
Trigram	No	No	Yes

PARSER: We implemented PARSER according to Perruchet and Vinter (1998) as well as personal communication with the first author. While a full description of PARSER is beyond the scope of the present article, it operates in much the same fashion as the Competitive Chunking model of Servan-Schreiber and Anderson (1990), but without building up a hierarchical network of chunks. Chunks are initially formed in PARSER through a stochastic process determining the size of percepts consisting of elementary units (in the present case, words). At each time step, chunks in the model's "lexicon" are affected by decay, with interference between partially overlapping chunks.

Perruchet (personal communication) advised us to retain the default values for the free parameters governing the threshold beyond which chunks shape percepts, the initial weight assigned to newly-segmented chunks, and weight added when existing chunks are reinforced by subsequent encounters (1.0, 1.0, and 0.5, respectively). Thus, the free parameters of primary interest for the present study were those governing decay and interference. We explored a range of values and adopted the one offering the best performance according to the gold standard for evaluating the models (described below).

While PARSER required no modifications to work with the shallow parsing task (merely the addition of a mechanism for recording its "percepts" as segmentations), Perruchet (personal

communication) declined to offer suggestions for how it might be applied to sequencing in the bag-of-words task. PARSER, according to Perruchet and Vinter (1998), was designed merely to build up an inventory of chunks rather than capture any sort of on-line usage of those chunks: "The issue addressed by PARSER is quite different, insofar as it concerns the *creation* of the lexicon (p. 252; emphasis in the original)." As such, we chose to focus the use of PARSER on the comprehension-related shallow parsing task only. Moreover, substantial changes to the model would be necessary in order to adapt it for use with the bag-of-words task.

Trigram baseline: To assess the usefulness of CBL and PARSER's variable-sized, recognition-based chunks as opposed to simpler sequential statistics tied to prediction, an additional alternate model was created which lacked a chunk inventory, relying instead on FTPs computed over stored *n*-grams. Since trigram models (second-order Markov models) are commonly used in computational linguistics as well as the field of machine learning (Manning & Schütze, 1999), we chose to focus on three-word sequences. This decision was further motivated by findings that trigram models are quite robust as language models, comparing favorably even to probabilistic context-free grammars. Our trigram model acquired statistics in an incremental, on-line fashion, in the style of CBL, while simultaneously processing utterances through the placement of chunk boundaries.

If the FTP between the first bigram and the final unigram of a trigram fell below the running average for the same statistic, a chunk boundary was inserted. For instance, as the model encountered Z after seeing the bigram XY, it would calculate the FTP for the trigram by normalizing the frequency count of the trigram XYZ by the count of the bigram XY, and comparing the result to the running average FTP for previously encountered trigrams (inserting a chunk boundary if the running average was greater). The start-of-utterance marker made it possible for the Trigram model to place a boundary between the first and second words of an utterance.

During production attempts, which were also incremental and on-line in the style of CBL, the

trigram model began constructing an utterance by choosing from the bag-of-words the word with the highest FTP, given the start-of-utterance marker (in other words, bigram statistics were used to select the first word). Each subsequent word was chosen according to trigram statistics, based on the two most recently placed words (or the initial word and the start-of-utterance marker, in the case of selecting the second word in an utterance). This meant the word with the highest FTP given the two preceding words was chosen at each time step.

Thus, the Trigram baseline model is purely prediction-based, by design. For instance, during production, it merely predicts the next word—because there is no recognition-based component, it does not work with entire chunks as recognition-based entities. The PARSER model baseline, by contrast, is purely recognition-based.

Simulation 1: Modeling Aspects of Child Comprehension and Production of English

In this section, we describe CBL simulations of child language learning and processing using English language corpora which capture interactions between children and their caretakers. We begin by describing the criteria used in selecting these corpora, followed by a description of the automated procedure used to prepare each corpus prior to its use as input in a simulation. Following this, we report the results of simulations for each corpus, comparing the performance of CBL to that of the two baseline models. For the sake of simplicity, we report performance for comprehension- and production-related tasks separately.

Corpus Descriptions and Preparation Procedure

In keeping with the key psychological features of the model, we initially sought to assess what could be

learned by CBL from the input available to individual children. We therefore selected developmental corpora involving single target children, rather than aggregating data across multiple corpora. From the English language sections of the CHILDES database (MacWhinney, 2000), we selected every corpus meeting the following criteria:

1) *Sufficient data* – In order to locate corpora that had sufficient diversity in terms of both vocabulary and syntactic constructions, we included only those corpora which contained at least 50,000 words.
2) *Dyadic* – Because we wished to model both comprehension and production for each child, we selected only corpora which featured a multiword child-to-adult utterance ratio of at least 1:10.
3) *Developmental* – As we sought to model the developmental progression of each child's language learning, we included only those corpora that spanned at least a 6-month period (in terms of the target child's age across the corpus).

The three criteria were met by corpora for 42 individual English-learning children (US: 25, UK: 17). For use in subsequent analyses, we collected, for each child, the age range (mean age of 1;11 at the beginnings of the corpora, 3;7 at the ends), number of months spanned by the corpus (mean: 20.6), total number of words in the corpus (mean: 183,388), number of child utterances (mean: 20,990), number of multiword child utterances (mean: 12,417), number of adult utterances (mean: 33,645), child mean length of utterance (MLU; the mean number of morphemes per utterance; mean: 3.17), and child mean number of words per utterance (mean: 2.6). For the full list of corpora, see Appendix D.

Corpus preparation: The corpora were submitted to an automated procedure whereby codes, tags, and punctuation marks were removed, leaving only speaker identifiers and the original sequence of words. To ensure that the input available to the model was representative of what children actually

receive, apostrophes were also removed from the corpora along with the other punctuation symbols. Thus, the contraction *it's* and the word *its*, for instance, were both represented orthographically as *its*, reflecting their identical phonological forms. This offered a naturalistic approach, considering developmental work indicating that children treat contractions as single words (cf. Tomasello, 2003).

Lines spanning tagged prosodic breaks, such as pauses (indicated in CHILDES by the (.) code), were broken into separate utterances, following research indicating that infants are sensitive to the suprasegmental properties of utterances, such as the acoustic correlates of clause boundaries (e.g., Hirsh-Pasek et al., 1987). Pauses due to hesitation (as indicated by the [/] code, etc.) were dealt with in the same manner. Finally, hash marks were added to the beginning of each line to signal the pause preceding each utterance.

Dense UK English corpus: The corpora in the CHILDES database typically represent a small percentage of the input a typical child might receive during the months spanned by the recording sessions. To examine subtle developmental trends with the model, a denser sample may be necessary. For this reason, we also tested the model using a dense corpus of child-directed speech which contains an estimated 8-10% of the target child's total productions (the Thomas corpus, originally known as the Brian corpus, which is now part of CHILDES; Maslen et al., 2004).

The dense corpus was submitted to the same automated procedure used to prepare the other CHILDES corpora. The prepared corpus spanned 36 months from age 2;0 to 5;0, featured 2,437,964 words, 225,848 child utterances, 114,120 multiword child utterances, 466,484 adult utterances, and an overall child MLU (in morphemes, as above) of 2.84.

Form class corpora: A considerable amount of work in computational linguistics has assumed that statistics computed over form classes are superior to item-based approaches for learning about structure (hence the widespread use of tagged corpora). This assumption is also present throughout the statistical learning literature (e.g., Thompson & Newport, 2007; Saffran, 2002), but is at odds with the

present model, which relies on statistics computed over concrete words and chunks rather than classes. To evaluate the usefulness of item-based chunking and statistics against those computed over word classes, we ran the model and its alternates on separate versions of each corpus, in which words were replaced by the names of their lexical categories. This process was automatically carried out by tagging each corpus using TreeTagger, a widely used, probabilistic part-of-speech tagger based on decision trees (Schmid, 1995). The tag set used by TreeTagger was reduced to the following 12 categories: noun, verb, adjective, numeral, adverb, determiner, pronoun, preposition, conjunction, interjection, infinitive marker, and proper name. Unknown words (e.g., transcribed babbling) were marked as such. As we removed the punctuation from each corpus as part of the preparation procedure, contractions were handled straightforwardly: contractions involving verbs were classed as verbs, while possessives were classed as nouns. Thus, contractions were classed according to the type of phrase they immediately appeared in (noun vs. verb phrases). This allowed us to avoid the use of a tokenizer (which would reflect an assumption that children represent contractions such as *don't* as two separate words), while being motivated by psychological considerations (e.g., a child may treat an utterance such as *that's the car* similarly to *see the car*; the verb-like aspect of the whole contraction takes precedence).

Grammaticality: The CHILDES corpora used are not marked for grammaticality. Thus, the present simulations do not distinguish between grammatical and ungrammatical utterances in scoring: as detailed below, production attempts are scored according to the same metric regardless of the grammaticality of the child's original utterance. While distinguishing between performance on a child's grammatical vs. ungrammatical utterances may be of interest in future work using corpora marked accordingly, such a project would involve a great deal of by-hand scoring by a large team of researchers and thus goes beyond the scope of the present work.

Evaluating Model Performance

Gold standard for testing comprehension performance of model and baselines: Shallow parsing: As the model approximated important aspects of comprehension by segmenting the incoming input into semantically related, phrase-like multiword units, we evaluated the model's comprehension performance—as well as that of the three baseline models—against the gold standard of a shallow parser. Shallow parsing is a widely used technique in the field of natural language processing, which aims to segment text into non-hierarchical (i.e., non-embedded) phrases which are labeled according to phrase type. As an example, take the sentence *the dog chased the cat.* A shallow parser would group the words together into noun and verb groups: *[NP the dog] [VP chased] [NP the cat]*. This choice of gold standard reflects the psychological motivation for the model; as observed by Sanford and Sturt (2002), shallow parsing identifies a subset of possible analyses for a sentence rather than giving the type of articulated analysis created by full syntactic parsers. This is in line with the previously discussed evidence for underspecification in sentence comprehension, as well as the shallow processing approach we adopt more generally, in which chunks of local information are used to arrive at a semantic interpretation of a sentence.

For each corpus, we generated a shallow parse for all utterances using the Illinois Chunker (Punyakanok & Roth, 2001), a widely used shallow parser based on constraint satisfaction with classifiers in a probabilistic framework. Phrase tags were then removed, leaving only the original sequence of words segmented via the phrase boundaries placed by the parser.

The model's on-line comprehension performance was scored according to two measures: *accuracy* and *completeness*, which are analogous to precision and recall, respectively. Each boundary marker placed by the model was scored as a *hit* if it corresponded to a boundary marker inserted by the shallow parser, and as a *false alarm* otherwise. Each boundary inserted by the shallow parser which was not placed by the model was scored as a *miss*. Thus, accuracy could be calculated as the proportion

of *hits* out of all boundaries placed by the model, *hits / (hits + false alarms)*, and completeness as the proportion of *hits* out of all boundaries placed by the shallow parser, *hits / (hits + misses)*. To avoid score inflation due to trivial factors, the model was only scored on utterance-internal boundaries (i.e., no boundary placement decisions were made at the beginnings or ends of utterances). Single-word utterances were excluded to avoid inflating the comprehension scores.

For purposes of scoring the model's comprehension performance on the form class corpora (in which individual items were replace by their lexical categories), the set of phrase boundaries placed by the shallow parser and used as the gold standard for scoring the original corpus was overlaid on the corresponding form class corpus. For instance, the utterance "[the dog] [chased] [the cat]," became "[DET N] [V] [DET N]" in the form class version, which therefore featured identical phrase boundary markers.

As an overall measure of comprehension performance for a given simulation, we relied on the F-score, which is widely used as a measure of performance in the fields of information retrieval and machine learning (e.g., van Rijsbergen, 1979). The F-measure combines both the precision (or *accuracy*, in the current case) and recall (*completeness*) of a test to compute a single score. We used the general F_β formula, which weights the completeness score according to β:

$$F_{\beta} = (1+\beta^2) * \left(\frac{accuracy * completeness}{(\beta^2 * accuracy) + completeness}\right) \quad (1)$$

In other words, the F_β metric attaches β times as much importance to completeness as to accuracy. In our case, β is the ratio of gold standard phrase boundaries to the total number of word pairs (the number of possible slots for boundary insertion) across a given corpus. The choice of the F_β metric reflects the need to control for score inflation stemming from trivial factors, such as over-segmentation

(e.g., due to data sparseness). As an example of this, consider a toy corpus which features phrase boundaries between exactly half of its word pairs. A model which heavily over-segments, placing phrase boundaries in every possible position, would receive a completeness score of 100%, and an accuracy score of 50%. By simply taking the harmonic mean of accuracy and completeness (what is known as the F_1 score), the model would receive an F-score of 66.67, despite its heavy over-segmentation. The F_β score, on the other hand, uses the number of word pairs straddling gold standard phrase boundaries to appropriately weight completeness in the calculation. For the previous example (where $\beta = .5$), this would yield an F_β-score of 55.56 (as opposed to the score 66.67 yielded by F_1) thereby reducing the impact of the perfect completeness score, which was achieved through trivial means (segmenting the corpus to the maximum extent).

Weighting accuracy more heavily than completeness in this way is also motivated by psychological considerations: phrases like *go to the shop* might be chunked as a single item by a child (as suggested by the results of Bannard & Matthews, 2008), or the model, whereas a shallow parser would segment it into three separate chunks: [*go*] [*to*] [*the shop*]. Therefore, the calculation also reflects the fact that accuracy, which reflects the model's ability to place boundaries that correspond to actual phrase boundaries (e.g., after *shop* or before *the* instead of between *the* and *shop*), may be more important than following the fine-grained chunking of a shallow parser (which penalizes the model through the completeness measure for not placing boundaries after *go* or *to* in a phrase like *go to the store*).

Gold standard for production: Child utterances. Each utterance produced by the model is evaluated against the corresponding child utterance in the original corpus, according to a simple all-or-nothing criterion: if the model's utterance matches the child utterance in its entirety, a score of 1 is assigned. In all other cases, a score of 0 is assigned, despite the degree of similarity between the model- and child-produced utterances. Thus, the overall percentage of correctly produced utterances provides

the *sentence production performance* for a given child/corpus. This represents a fairly conservative measure, as the model may produce sentences that are grammatical but nevertheless fail to match the target utterance. For example, the model may produce a sentence such as *the cat chased the dog* when the target sentence is *the dog chased the cat*. In such instances, the model receives a score of 0, due to the lack of principled and efficient way of automatically evaluating mismatching utterances that are nevertheless grammatical.

Parameter selection for PARSER: Following communication with the model's creator (Perruchet, personal communication), we adjusted the interference and decay parameters along a wide range, maintaining their separation by a factor of ten (following the settings used by Perruchet & Vinter, 1998), and selected the one offering the best combination of accuracy and completeness. At higher settings (e.g., Decay: 0.001, Interference: 0.0001), the model heavily over-segmented, placing boundaries between 90% of words. At settings 0.0001 and 0.00001 (for decay and interference, respectively), the model saw substantial improvements in accuracy while segmenting at a rate comparable to the CBL model. Decreasing the parameters by a further factor of ten lead to slight drops in accuracy and completeness. Thus, for the natural language simulations, we adopted settings of 0.0001 and 0.00001 for decay and interference, respectively.

Results and Discussion: Simulating Aspects of Comprehension and Production of English

Shallow parsing performance. Across all 43 single-child English corpora, CBL attained a mean F-score of 75.4, while the PARSER model attained a mean F-score of 66.1. The Trigram model had a mean F-score of 65.9. Comprehension performance for each model is shown in Figure 3. As can be seen, the CBL model not only outperformed its baselines, but yielded a tighter, more uniform distribution of scores across the corpora.

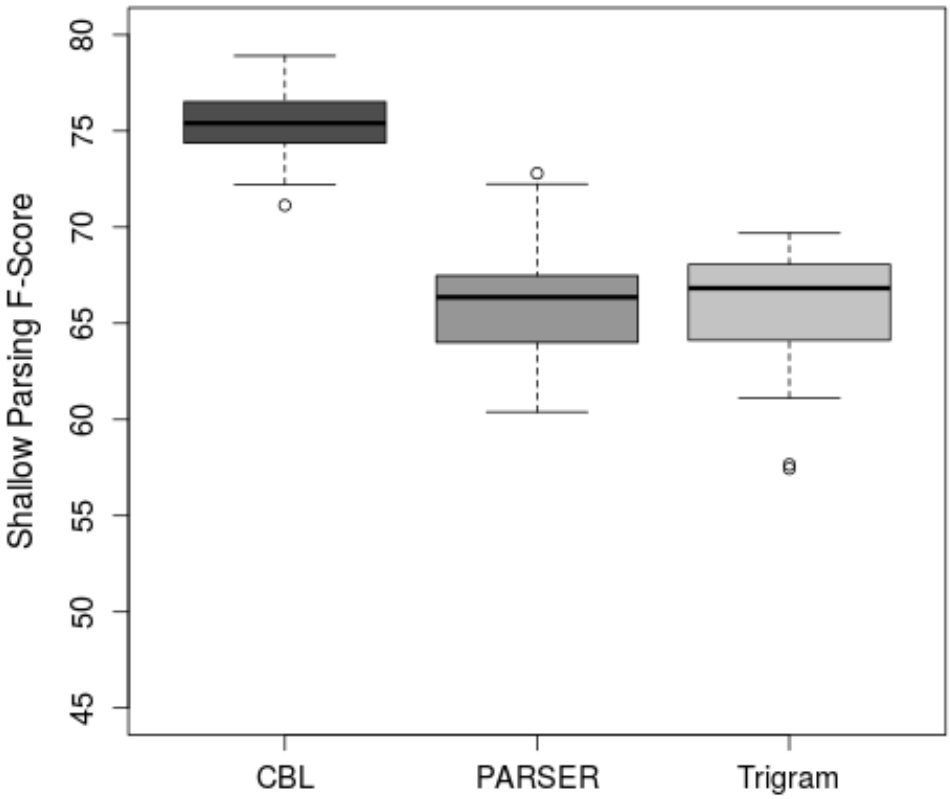


Fig. 3: Boxplots depicting English shallow parsing F-scores for the CBL model and its baselines. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

The F-scores for the model and its baselines were logit-transformed[7] and submitted to a repeated-measures ANOVA including the factor *Model* (3: CBL vs. PARSER vs. Trigram) with *Child Corpus* as a random factor. This yielded a significant main effect of *Model* [$F(2,84) = 643.3$, $p < 0.0001$], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [$t(42)=35.9$, $p<0.0001$] and Trigram [$t(42)=28.3$, $p<0.0001$] models, with no significant difference in

7 As the scores necessarily have both floors and ceilings, and represent proportional data, a logit transformation was applied prior to analysis in order to fit the assumptions of the test.

means between PARSER and the Trigram model [$t(42)=0.49$, $p=0.63$].

In line with the developmental motivation for the model, we also examined accuracy rates independently. Across the 43 child corpora, CBL achieved a mean accuracy rate of 76.4%, while PARSER attained a mean accuracy of 65.2% and the Trigram model reached a mean accuracy rate of 65.8%. The same general pattern was seen for completeness: CBL achieved a mean completeness of 73.8%, while the PARSER attained a mean completeness of 68.7% and the Trigram model reached a mean completeness rate of 66.5%. Accuracy and completeness scores are described more fully in Appendix A.

Thus, the best combination of accuracy and completeness (as measured by the F-score), as well as the best accuracy and completeness overall, was achieved by CBL's statistically-based chunking for the English child corpora. CBL was able to approximate the performance of a shallow parser through a combination of recognition- and statistically-based processing in an on-line, incremental fashion starting with a single distributional cue. This result is encouraging, as shallow parsing is regarded as a nontrivial problem in the field of natural language processing (e.g., Hammerton, Osborne, Armstrong, & Daelemans, 2002).

In addition to highlighting the wealth of distributional information in the input, these results suggest that purely item-based information may be far more useful to early learners than has been assumed previously; in addition to providing the basis for discovering useful multiword sequences (which may later be abstracted over, as proposed by usage-based approaches more generally; e.g., Tomasello, 2003), statistical information tied to concrete items can help uncover chunks of local information necessary to interpret sentences (“phrase structure,” in most approaches), as demonstrated by the present model.

It is also worth noting that the CBL model tends to discover chunks at a coarser level of granularity than the shallow parser used as a gold standard, as reflected by the difference in accuracy

and completeness scores. As noted above, phrases like *go to the shop* form useful chunks for a child (as suggested by Bannard & Matthews, 2008), whereas the gold standard posits three separate chunks: [*go*] [*to*] [*the shop*]. Therefore, accuracy, which measures the model's ability to place boundaries corresponding to actual phrase boundaries, may be more useful for our purposes than completeness, given the fine-grained chunking of a shallow parser (as completeness would involve penalties for not placing boundaries after *go* or *to* in a phrase like *go to the store*).

CBL and the Trigram model may have outperformed PARSER in part because of the latter model's over-reliance on raw frequency of occurrence. For instance, CBL can identify high TPs between items which have occurred with very low frequency in the corpus: the relative, rather than absolute, frequency of the two items is stressed. While PARSER is indirectly sensitive to TPs, via its decay and interference parameters, the use of these parameters along with randomly determined percept sizes may requires more exposure.

CBL also has the additional advantage of being directly sensitive to *background rates* (cf. Ramscar, Dye, & McCauley, 2013). Words that occur extremely often in a variety of contexts have high background rates, which mean they are less informative about the items preceding them (or following them, in the case of the Trigram model). Conditional probabilities directly reflect this. PARSER is only indirectly sensitive to background rates, through its interference feature: items that occur often as parts of larger chunks will lead to decreases in the strength of chunks featuring the same item. However, the impact of the decay parameter on less-frequent chunks may still lead to an overemphasis on items with high background rates.

For these reasons, PARSER may ultimately be best suited to working with small, artificial languages which involve fairly uniform frequency distributions over items, such as those featured in studies to which it has previously been applied (e.g., Perruchet et al., 2002; Saffran et al., 1996).

Development of the chunkatory. The development of the model's knowledge over the course of a simulation, independently of its scored performance, may offer potential predictions on which to base future psycholinguistic work. To provide a snapshot of the types of sequences chunked by the model, we provide in Appendix B a breakdown of the most highly activated chunks in the model's inventory at three separate points in development. Table B1 shows this for individual items, while Table B2 shows this after conversion of each chunk into lexical categories (e.g., *the dog* and *the cat* would both be counted as *determiner noun* and both contribute to the frequency count thereof).

Briefly, as can be seen in Appendix B, the most highly activated chunks cover a range of usage contexts, extend beyond just noun- and verb-phrases, and have significance for connecting the model to existing work on the role of chunks in language development. For instance, units like *I think* are highly relevant to previous work on the role of chunks in the acquisition of finite complement structures (e.g., Brandt, Lieven, & Tomasello, 2010; Diessel & Tomasello, 2001), while units covering wh- formulae (e.g., *what's this*) make contact with work on question development (e.g., Ambridge, Rowland, & Pine, 2008; Rowland, 2007).[8] In Simulation 2, below, we discuss the model's ability to directly simulate data from developmental psycholinguistic studies.

To examine the development of the chunkatory more generally, we tracked the percentage of stored chunks which consisted of multiple words, for chunk types as well as chunk types weighted by their frequency counts in the chunkatory (akin to chunk tokens). Figure 4a shows the percentage of multiword chunks in the chunkatory as it develops, for each of the 43 child corpora. As can be seen, the percentage of multiword chunk types increased logarithmically over the course of the simulations, leveling off below 90%. The first data point fell above 50% for all child corpora. When we weighted individual chunk types by their strength (frequency counts) in the chunkatory, however, we found

[8] We thank an anonymous reviewer for suggesting that we highlight these connections with items appearing prominently in the chunk inventory.

higher percentages for single-word chunks, as shown in Figure 4b.The percentage of multiword chunk types, when weighted by frequency, began well below 50% at the first data point for all corpora, with percentages for many of the child simulations dipping within the first 20,000 utterances before rising sharply and then climbing more steadily.

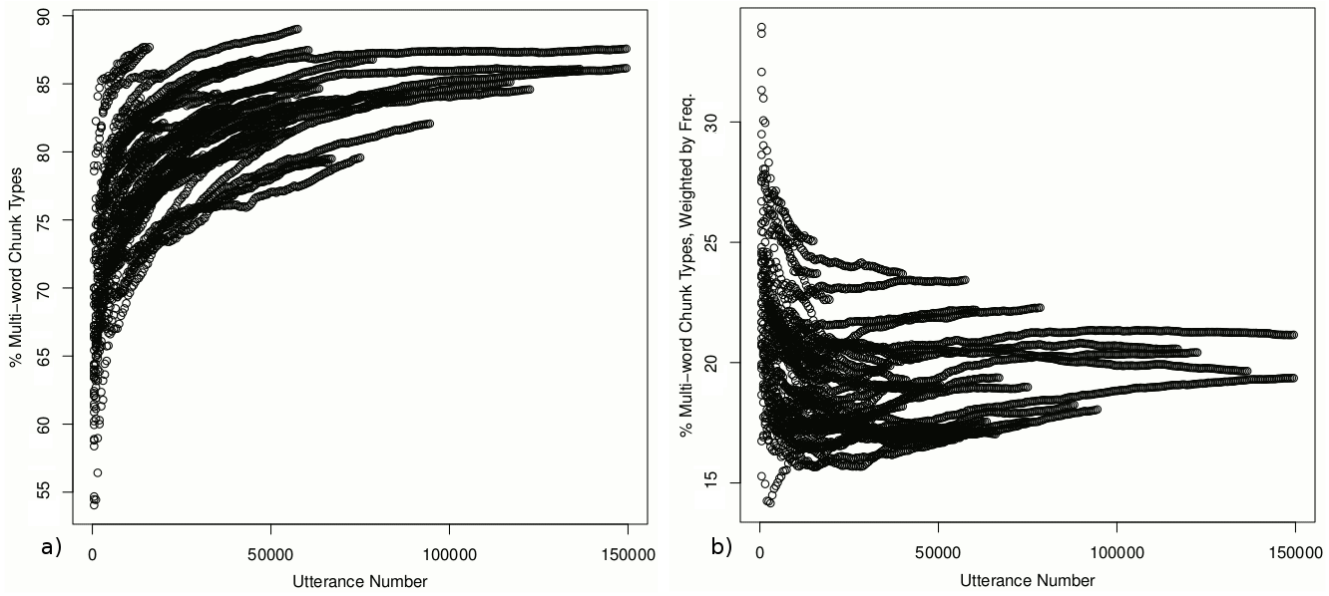
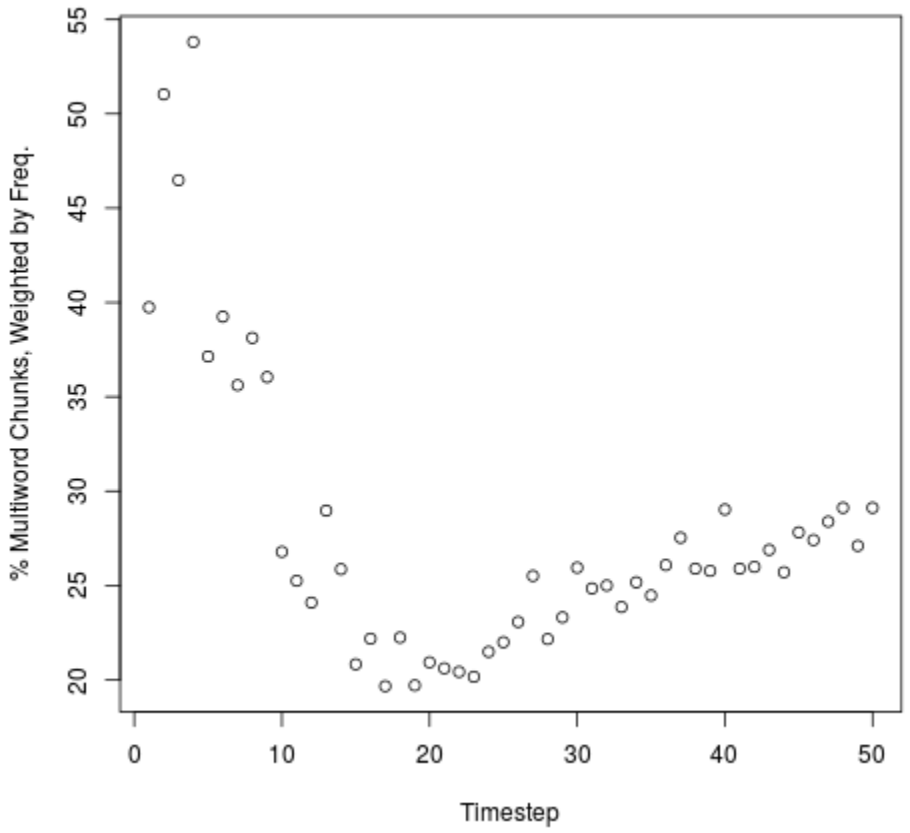


Fig. 4: a) Development of the English chunkatory by percentage of multiword types; b) Development of the English chunkatory by percentage of multiword types weighted by frequency of use.

The dense corpus (Thomas) shows the same pattern. To look more closely at the makeup of chunks we might expect the child to actively use in production, we calculated the percentage of multiword chunk types, weighted by frequency, which were actively used by the model during the bag-of-words task. This is depicted in Figure 5 for the dense corpus (Thomas).

Fig. 5: Percentage of multiword types weighted by frequency of use in the production task for the dense corpus of



English (Thomas). Each time step represents the mean percentage across 2,000 child utterances.

As can be seen, the prominence of multiword chunks in the chunkatory for the Thomas simulation mirrors the general pattern illustrated by Figure 4b, dipping early before rising once more. However, because the dense corpus extends well past the other corpora in terms of length, we were able to look at a more complete trajectory, which included a sharp dip followed by a more subtle increase which spanned the remainder of the simulation. This pattern derives from: 1) an initial period in which newly-encountered words are naturally chunked with preceding material owing to transition probabilities at or approaching 1, yielding large chunks; 2) a period in which the model gains more exposure to words and increasingly discovers more fine-grained units, rapidly reducing the average chunk size as a result; and 3) a more gradual increase in the average chunk size as the model gains

enough exposure to combine its knowledge of chunks through its on-line "recognition-based prediction" mechanism (which assists in chunk formation based on previously learned chunks, as described above).

The U-shaped curve exhibited by the model mirrors a common developmental pattern which has been tied to several aspects of language learning, including phonological development (Stemberger, Bernhardt, & Johnson, 1999), morphological development (Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992), relative clause comprehension (Gagliardi, Mease, & Lidz, submitted), and verb usage (Alishahi & Stevenson, 2008; Bowerman, 1982). The model's behavior therefore points to the prediction that children's reliance on multiword chunks may shift in similar ways to that of the model, and that this may have some bearing on U-shaped trajectories in other areas, such as morphological development. For instance, Arnon and Clark (2011) found that over-regularization errors were less likely when irregular plurals were produced in the context of a lexically-specific frame; the facilitatory role played by chunks in this area (and others) may wax and wane with the "degree of chunkedness" of the child's linguistic representations, consistent with preliminary findings demonstrating a U-shaped pattern for this over-regularization effect across children of different ages (Arnon, personal communication). The overall trajectory of CBL's chunk development therefore leads to a concrete prediction, to be tested in future developmental psycholinguistic work.

In summary, multiword chunks ultimately grow in importance to the model over the course of a simulation, both in terms of types and in terms of tokens. In light of psycholinguistic work with adults (Arnon & Snider, 2010; Bannard & Ramscar, 2007), this leads us to predict that children do not merely "start big" by relying on larger multiword sequences which break down over time, leaving single words; rather, the child's memory-based processing is dynamic, and the degree to which representations of linguistic material are tied to multiword sequences ultimately grows in importance over time. The model's U-shaped reliance on weighted multiword chunks also leads us to propose that children may go

through periods where new knowledge of the properties of single words may lead to a decreased reliance on multiword sequences, only to be followed by a renewed reliance on chunked representations.

***Class- vs. item-based comprehension performance*.** As discussed above, most generative approaches as well as certain trends within the statistical learning literature (cf. Thompson & Newport, 2007) have assumed that language learning is tied to word classes. For this reason, we re-ran the 43 simulations reported above, using the form class corpora (see the corpus preparation above for a description of how words were converted to form classes).

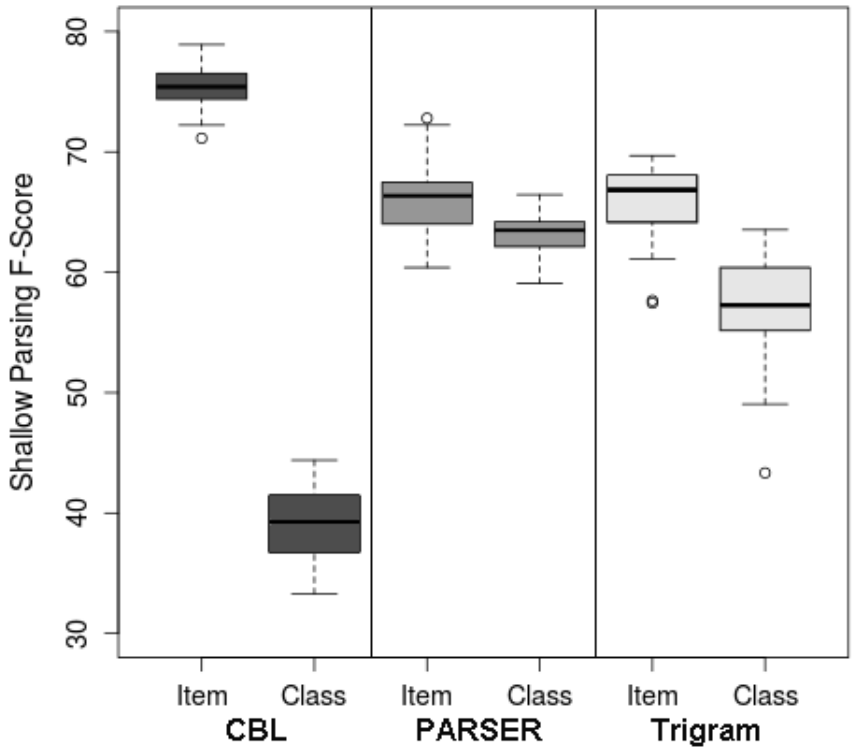
Because PARSER is sensitive to overall number of unit types it is exposed to, we found that the highest parameter setting we tested for natural language (0.01 for decay, and 0.001 for interference) provided the best trade-off between accuracy and completeness when working with form classes.

Class- versus item-based performance for the model and its baselines is depicted in Figure 6. Performance for CBL was considerably worse when working with class-based statistics, with a sharp decrease in the mean F-score (from 75.4 to 39). For PARSER there was a far less drastic decrease in performance (from a mean F-score of 66.1 to 63.1). The Trigram baseline also fared worse under class-based statistics, though with less dramatic decreases in performance. The mean F-score dropped from 65.9 to 57.2.

In the case of the CBL model, the lower comprehension performance when working with class statistics was driven both by a drop in accuracy as well as a more drastic drop in completeness scores, the latter owing partly to the use of the chunkatory in phrase segmentation; the relatively small number of possible class combinations in a sequence lead to the automatic grouping of items together (based on the use of the chunkatory) with increasing frequency throughout the models' pass through a corpus. As more combinations exceeded the average TP threshold, the models placed progressively fewer phrase boundaries. PARSER, however, saw a slight increase in accuracy accompanied by a decrease in

completeness. As PARSER was designed for use with small item sets, as discussed above, further experimentation with the parameter settings of PARSER may be necessary in order to improve performance on the form-class simulations.

Fig. 6: Boxplots depicting English comprehension performance (F-scores) for the CBL model and its baselines,



comparing item-vs. class-based simulations. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within the 1.5 IQR of the quartiles, while dots depict outliers.

We evaluated the effects of learning from class-based information using a two-way ANOVA with factors *Statistic Type* (2: item- vs. class-based) and *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor, over logit-transformed F-scores. This yielded main effects of

Statistic Type [$F(1,42) = 1562$, $p<0.0001$], confirming stronger performance for item-based models, and of *Model* [$F(2,84)=171.4$, $p<0.0001$], with an interaction between *Statistic Type* and *Model* [$F(2,84)=25.5$, $p<0.0001$], due to a more drastic drop in performance for the CBL model relative to the baselines when working with classes.

Thus, a reliance on word classes did not improve the performance of the model or its baselines; instead, knowledge of classes lead to a decrease in performance, which was considerably more drastic for CBL. This result makes close contact with item-based approaches more generally (e.g., Tomasello, 2003), suggesting that successful language learning can begin without the sort of abstract syntactic categories that much previous psycholinguistic and computational work has focused upon. This also runs counter to claims made in the statistical learning literature that children and adults use transitional probabilities to segment phrases by calculating statistics over word classes rather than concrete items (e.g., Saffran, 2002; Thompson & Newport, 2007). Indeed, we have shown elsewhere (McCauley & Christiansen, 2011) that comprehension through item-based learning in our model captures subject performance in one such study (Saffran, 2002) better than class-based learning.

Production performance. Across all 43 single-child corpora, CBL achieved a mean sentence production performance of 58.5%, while the Trigram model achieved a sentence production performance score of 45.0%. Recall that PARSER was not compatible with the production task (for reasons discussed in the Methods section above). The distributions of the scores for each model are depicted in Figure 7. As can be seen, the overall pattern of results was similar to that seen with comprehension, with CBL achieving the highest mean score.

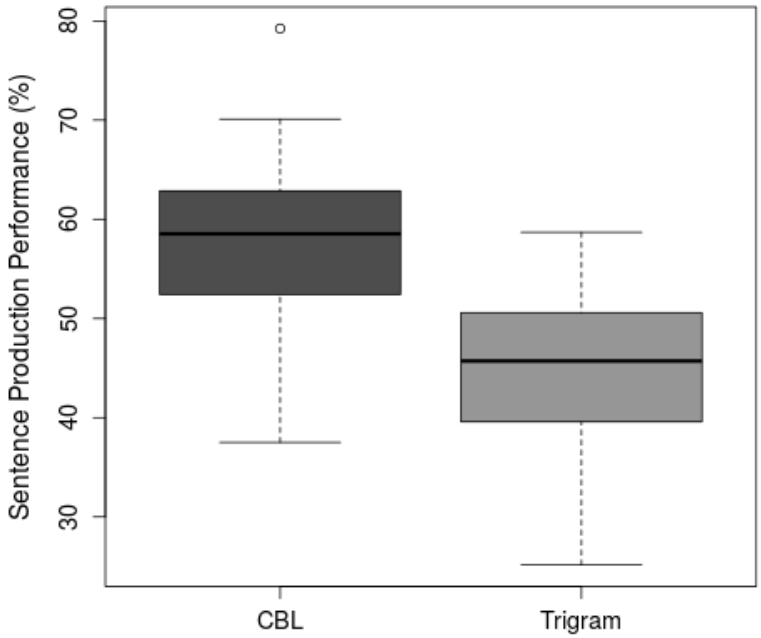


Fig. 7: Boxplots depicting English Sentence Production Performance scores for the CBL model and the Trigram baseline. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

A repeated-measures ANOVA including the factor *Model* (2: CBL vs. Trigram), with *Child Corpus* as a random factor, yielded a significant effect of *Model* [$F(1,42) = 514.9$, $p < 0.0001$], indicating better performance for CBL.

Thus, CBL exhibited clear advantages over the baseline. The advantage of stored-chunks (which do not have to be sequenced, in and of themselves) is clear in these results. What is less clear is that BTPs may offer an advantage over FTPs when there is a limited, specified set of possible items that can follow the most recently placed item in a sequence (such as a bag-of-words, in the present instance). The FTP-based Trigram model simply selects, at each time step, the item combination with the highest frequency, since the frequency of every possible sequence is normalized by the preceding

item, which is fixed, as it has already been produced. The CBL model, however, through the use of BTPs, is sensitive to the background rate (discussed above; cf. Ramscar et al., 2013) of the candidate items: items that occur more often in contexts other than the present one will not be selected.

Production summary and discussion. CBL not only outperformed its baselines in reproducing child utterances, but was able to produce the majority of the target utterances it encountered, with a mean score of nearly 60% based on our conservative all-or-nothing measure of production performance. This not only underscores the usefulness of chunks and simple statistics such as BTPs, but serves to demonstrate that the same sources of information can be useful for learning about structure at multiple levels: a single distributional statistic (BTP) can be used to segment words when calculated over syllables (e.g., Pelucchi et al., 2009), to discover multiword sequences (or perhaps even "phrase structure") when calculated over words (demonstrated by the present model), and to construct utterances when calculated over stored multiword chunks themselves (demonstrated in the current section).

Despite this success, the model nevertheless failed to account for 40% of the child utterances, a significant proportion, and yielded a less dramatic advantage over its baseline than was the case with comprehension. This pattern of results, when considered alongside the idea of shallow processing as a central feature of a child's language comprehension, has immediate implications for the comprehension/production asymmetry in children, insofar as it stems from differing task demands. Through shallow processing of the sort captured by the model, a child can give the appearance of having utilized a construction (such as a transitive construction in canonical word order) in comprehension while still lacking the sequential knowledge to use it in production. This is especially true if one considers specific aspects of shallow processing in adults, as well as its ubiquitous nature in language comprehension more generally. Take, for instance, passive sentences. Ferreira (2003) found that when adults were exposed to anomalous sentences using passive constructions (*"The dog was*

bitten by the man") many readers utilized global pragmatic information rather than the passive construction to identify agents and patients of actions, and gave higher plausibility ratings than for the same content given in active voice (i.e., they interpreted the passive sentence as meaning that the dog bit the man). If even adults rely as much on local information, pragmatic expectations, and background world knowledge to interpret sentences as on the actual constructions used in the sentence, it seems likely that children also will depend on such information, giving the appearance of having fully exploited a grammatical construction when in actuality they merely interpreted the utterance at a more superficial level.

Under such a view, children may understand specific utterances utilizing passive voice without having mastered the passive construction. By contrast, to correctly sequence the words (or word chunks) in a passively voiced sentence, the child would necessarily need to have substantial knowledge of the passive construction schema (such as *PATIENT is ACTION by AGENT*) as well as knowledge of the pragmatic motivations for using it (for a model which learns to produce sentences using such semantic role information, see Chang, Dell, & Bock, 2006). Thus, shallow processing allows a child to make a great deal of progress towards understanding language input merely on the basis of an ability to chunk parts of utterances and form associations between those chunks and concrete parts of the world, as well as event schemas or scenarios; it is in the sequencing of those chunks that the problem of production becomes more difficult than that of comprehension. This idea is explored further using the CBL model by Chater et al. (2016).

The relationship between comprehension- and production-related processes in the model is especially relevant in light of current theoretical perspectives that view comprehension as fundamentally tied to prediction in ways that are mediated by production itself (e.g., Martin, Branzi, & Bar, 2018; Pickering & Gambi, 2018). Future work should aim to more fully integrate production- and comprehension-related processes in the model in order to explore such perspectives in the context of

child language development.

Interim summary: Learning English. We have shown that the CBL model is able to approximate shallow parsing through the incremental, on-line discovery and processing of multiword chunks, using simple statistics computed over local information. The model is able to use the same chunks and statistics to produce utterances in an incremental fashion, capturing a considerable part of children's early linguistic behavior in the process. This is achieved through item-based learning, without recourse to abstract categorical information such as that of word classes. When the model learns class-based statistics, its ability to segment useful chunks is impaired. Furthermore, the development of the model's chunk inventory offers the novel prediction that subtle shifts in the "degree of chunkedness" of children's linguistic units may impact on other areas of language development. Finally, the model, which combines chunking with statistical cues, compares favorably to exclusively recognition-based and exclusively prediction-based baselines.

Simulation 2: Modeling the Development of Complex Sentence Processing Abilities

Whereas the previous simulations examined the ability of CBL to discover building blocks for language learning, in the present section we investigate the psychological validity of these building blocks. Previous modeling work has demonstrated the ability of CBL to fit developmental psycholinguistic data related to children's chunk-sensitivity and morphological development (cf. McCauley & Christiansen, 2014a), while work by other researchers has demonstrated the psychological validity of CBL's chunk discovery using reaction time patterns (Grimm et al., 2017).

In the present section, we report simulations of empirical data covering children's ability to process complex sentence types (Diessel & Tomasello, 2005). Usage-based approaches predict that stored chunks play an integral role in the development of complex grammatical abilities (e.g.,

Christiansen & Chater, 2016a; Tomasello, 2003), which have been argued to emerge from abstraction over multiword sequences (e.g., Goldberg, 2006; for models, see Kolodny et al., 2015; Solan et al., 2005).

Nevertheless, there is strong evidence of a role for concrete multiword chunks in adult processing of grammatically complex sentences, such those featuring embedded relative clauses (e.g., Reali & Christiansen, 2007), which in turn suggests that children's ability to comprehend and produce complex sentences should be influenced by the same type information. If this holds true, and if CBL provides a reasonable approximation of children's discovery and use of chunks, the model should be able to offer some insight into the development of complex grammatical abilities, despite its lack of abstract grammatical knowledge. In order to test this notion, we used CBL to model children's ability to produce different relative clause types (Diessel and Tomasello, 2005), as a great deal of previous developmental work on grammatically complex sentences has focused on relative clause constructions (see Christiansen & Chater, 2016a, for a review).

This particular study was chosen because its stimuli were designed to reflect the types of relative constructions children actually produce in spontaneous speech (specifically, those that attach to either the predicate nominal of a copular clause, or to an isolated head noun; Diessel, 2004; Diessel & Tomasello, 2000). Prior to this study, developmental work on relative clauses focused mainly on sentence types which children rarely produce spontaneously, and which therefore may not adequately reflect children's grammatical knowledge (e.g., Hamburger & Crain, 1982; Keenan & Hawkins, 1987; Tavakolian, 1977). Of further importance is the study's focus on children's production abilities as opposed to just comprehension; because the stimuli consisted of whole sentences, this allowed us to model child performance using the entire model architecture (comprehension as well as production).

Using a repetition paradigm, Diessel and Tomasello exposed a group of UK English-speaking

children (mean age: 4;7) to sentences featuring one of six relative clause types: *subject relatives featuring intransitive verbs (S), subject relatives featuring transitive verbs (A), direct-object relatives (P), indirect-object relatives (IO), oblique relatives (OBL), and genitive relatives* (GEN). An example of each relative clause type is shown in Table 2. Following exposure to a sentence, the child was prompted to repeat it to the experimenter. The authors found that children's ability to reproduce the relative clauses closely mirrored the similarity of each clause type to simple non-embedded sentences (with the greatest accuracy for subject relatives).

Table 2

Relative Clause Types from Diessel and Tomasello (2005)

Type	**Example**
S	There's the boy who played in the garden yesterday.
A	That's the man who saw Peter on the bus this morning.
P	That's the girl who the boy teased at school this morning.
IO	There's the girl who Peter borrowed a football from.
OBL	That's the dog that the cat ran away from this morning.
GEN	That's the woman whose cat caught a mouse yesterday.

Method. We began by exposing CBL to a corpus of UK English. As the original English study was conducted in Manchester, we focused on the dense Thomas corpus (which was recorded in Manchester; Maslen et al., 2004). Following exposure to the corpus, CBL was presented with the same test sentences heard by children in the original study. Immediately following comprehension on a given test sentence, the model simulated a repetition trial by attempting to produce the utterance (using the bag-of-words task in an identical manner to the child utterances in our original natural language simulations). If the utterance produced by the model matched the target utterance in its entirety, a score of 1 was assigned; otherwise, a score of 0 was assigned.

To order the test items, we used the same randomization procedure as was used in the original study: items were organized into four consecutive blocks of six randomly chosen sentences, with the constraint that each block included one sentence from each condition (Diessel, personal communication). This randomization allowed for small individual differences to arise between simulations (21 in total; a different randomization/simulation pair for each child in the original study).

Results and discussion. The children in the original study achieved the following correct response rates, as shown in Figure 8: 82.7% (S), 59.5% (A), 40.5% (P), 31% (IO), 31.5% (OBL), and 2.5% (Gen). As also shown in Figure 8, correct response rates for the model were 77.4% (S), 48.8% (A), 75% (P), 39.3% (IO), 34.5% (OBL), and 16.7% (GEN). As can be seen, the model followed the same general pattern as the children in the original study, with the exception of its performance on P-Relatives, which was almost as high as its performance on S-Relatives.

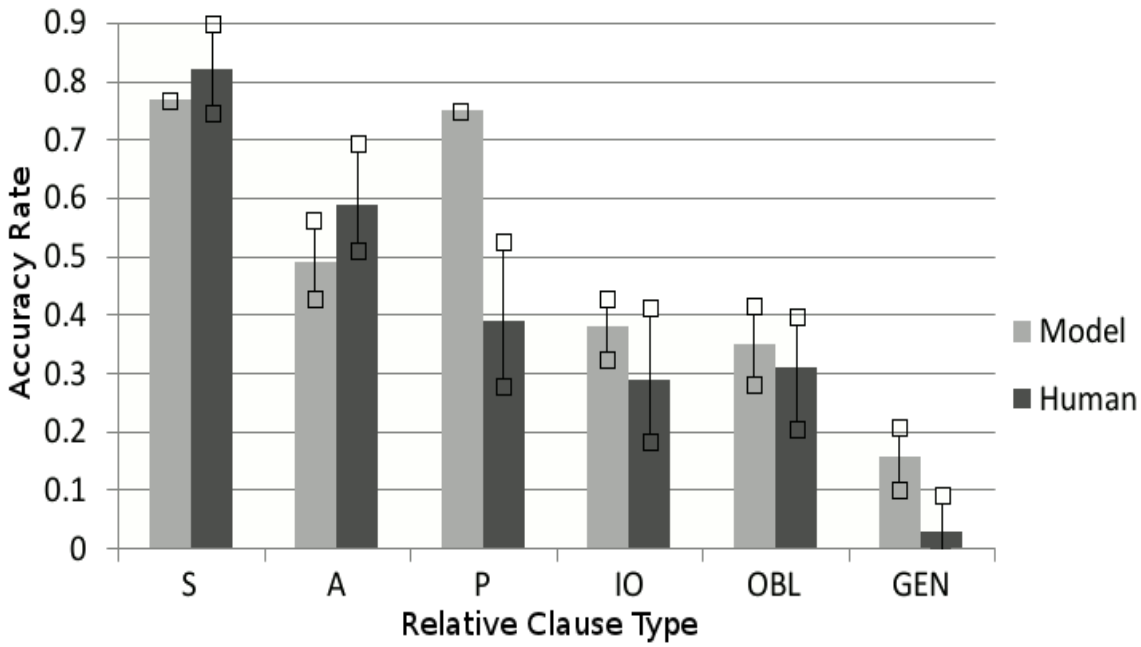


Fig. 8: Mean correct response rates for CBL model and the child subjects in Diessel & Tomasello (2005). Error bars denote 2x standard error.

That the model was able to mirror the child repetition performance for most of the clause types is unexpected, considering its complete lack of semantic/pragmatic information or distributional information spanning non-adjacent chunks. Previous connectionist modeling work has successfully captured data from the Diessel and Tomasello study by incorporating structured meaning representations (Fitz & Chang, 2008). While this work is somewhat limited in that it involves the use of hand-generated datasets as input rather than child-directed speech, it nonetheless suggests a crucial role for meaning in explaining processing differences across relative clause types, a role which is emphasized in subsequent modeling work on subject-auxiliary inversion in question formation (Fitz & Chang, 2017). Following such work, we view the incorporation of meaning as a key future challenge in the extension of distributional models such as CBL (this challenge is further discussed below; see also McCauley & Christiansen, 2014b).

Despite the model's decent fit to the child data for 5 of the 6 relative clause types, its over-performance on the P-Relatives serves as a reminder of the limits of a purely distributional approach; semantics obviously plays a role not only in children's on-line chunking of test utterances upon first exposure (corresponding to the comprehension side of the model), but their incremental production during repetition attempts (corresponding to the sequencing stage of production in the model), and their recall of chunks throughout both (corresponding to recognition-based processing during comprehension and the retrieval stage of production in the model). As the model receives no input related to semantic roles, it received no information on the *patient* role of the main clause subject within the P-Relatives, and hence no interference from its deviation from the *agent-action-patient* sequence most commonly encountered in simple non-embedded sentences. Instead, the model relied on purely item-based similarity or dissimilarity to sentences in the child-directed speech it initially learned from. Moreover,

the model's high performance on P-Relatives might also stem in part from the nature of its input: P-Relatives may be even more common in English child-directed speech than A- and S-Relatives. In an analysis by Diessel (2004), 56.8% of all relative clauses produced by four English-speaking parents were P-Relatives, while 35.6% were S- or A-relatives, 7.6% were OBL-Relatives, whereas IO- and GEN-Relatives did not occur.

Diessel and Tomasello (2005), in accord with usage-based approaches, propose that young children's ability to produce relative clauses depends on the degree of similarity between the type of relative clause and simple non-embedded sentences of the sort encountered most often in child-directed speech (an idea which has received considerable empirical support in recent years; e.g., Brandt, Diessel, & Tomasello, 2008; see Christiansen & Chater, 2016a, for a review). This stands in contrast to the hypothesis that the distance between filler and gap determines processing difficulty (Wanner & Maratsos, 1978), which initially sought to explain the well-documented phenomenon of greater processing difficulty for object relative as opposed to subject relative clauses (as demonstrated in Dutch, English, and French; Wanner & Maratsos; 1978, Frauenfelder, Segui, & Mehler, 1980; Holmes & O'Regan, 1981; Ford, 1983; Frazier, 1985; King & Just, 1991; Cohen & Mehler, 1996; though see also Brandt, Kidd, Lieven, & Tomasello, 2009). Under this view, object relatives cause more difficulties than subject relatives because they feature a greater distance between filler and gap, and the filler must be retained in working memory until the gap is encountered. The distance between filler and gap has also been hypothesized to play a role in the ease of acquisition of relative clauses, favoring relative clauses with a short distance between filler and gap (e.g., de Villiers et al., 1979; Clancy, Lee, & Zoh, 1986; Keenan & Hawkins, 1987).

In CBL, both comprehension and production rely on statistics computed over adjacent chunks; the model has no "working memory," and thus no sensitivity to the distance between filler and gap in

relative clause constructions. Nevertheless, we observe better performance on the S-Relatives (for which the distance between filler and gap is the smallest) than other relative clause types. At the same time, CBL fits the pattern of better performance on the S- than A-Relatives exhibited by the children in the original study; as Diessel and Tomasello note, the distance between filler and gap cannot account for this result, as the distance is the same in both relative clause types. Furthermore, we observed the worst performance with GEN-Relatives (again, a pattern exhibited by the child subjects), despite the small distance between filler and gap for these sentences. Our results therefore have a direct bearing on the filler-gap hypothesis, beyond merely reinforcing Diessel and Tomasello's findings, suggesting that children's item-specific knowledge may play a greater role in relative clause processing than working memory constraints (see Christiansen & Chater, 2016a; MacDonald & Christiansen, 2002; McCauley & Christiansen, 2015, for extensions of this perspective to individual differences in adult relative clause processing).

The results of this simulation also allow us to derive a novel prediction from the model: the very factor driving the model's performance, item-based statistics computed over adjacent chunks, may well be a factor in the apparent ease with which children learn to produce certain kinds of sentences while encountering difficulties in learning to produce others. Indeed, previous evidence from children's elicited question formation indicates a role for such surface distributional statistics (Ambridge, Rowland, & Pine, 2008). This prediction can be tested further with a simple repetition paradigm such as that used by Diessel and Tomasello, using stimuli which systematically pit adjacent chunk statistics against statistics derived from large corpora of child and child-directed speech, in such a way that local information conflicts with the global properties of coherent target utterances.

Modeling Child Comprehension and Production across a Typologically Diverse

Array of Natural Languages

We have shown that CBL can capture a considerable part of children's early linguistic behavior, in addition to making close contact with developmental psycholinguistic data from a key study on children's item-based distributional learning. Nevertheless, these findings—like most of the psycholinguistic findings forming the basis for the model—are based entirely on the use of English data; the computational approach we have adopted may not actually characterize aspects of learning held in common by learners of typologically different languages. In the next series of simulations, we explore the question of whether the model can extend beyond English to cover a typologically diverse set of languages.

The goal of attaining broad, cross-linguistic coverage extends beyond merely building support for the model; we aim to address certain limitations of the psycholinguistic literature. For instance, a potential problem with the view of multiword chunks as an important feature of language use is that most of the directly supporting psycholinguistic evidence has been gathered from English-speaking subjects. Importantly, English is an *analytic* language; it has a low ratio of words to morphemes, relative to *synthetic* languages, which have higher ratios due to the many ways in which morphemes can be combined into words. What may apply to arguments about unit size in the learning of analytic languages (such as Mandarin or English) may not apply to the learning of synthetic languages (such as Tamil or Polish), and vice versa. It is therefore essential to test the predictions of both CBL and previous empirical work with English-speaking subjects by modeling chunk-based learning across a typologically diverse set of languages. The breadth of material in the CHILDES database (MacWhinney, 2000) makes it possible to test the model on a typologically diverse array of languages.

Following a description of the corpora used to simulate learning cross-linguistically, we report comprehension performance for the languages for which an automated scoring method was available.

We then report sentence production performance for 28 additional languages.

Corpus Selection and Preparation

Corpora were selected from the CHILDES database (MacWhinney, 2000), and covered a typologically diverse set of languages, representing 15 genera from 9 different language families (Haspelmath, Dryer, Gil, & Comrie, 2005). As with the English simulations, we sought to assess what could be learned by CBL from the input available to individual children. We therefore selected, once more, developmental corpora involving single target children, rather than aggregating data across multiple corpora. However, due to the limited availability and size of corpora representing several of the languages in the CHILDES database (MacWhinney, 2000), we relaxed our criteria somewhat. Thus, we used corpora that met the following criteria:

1) *Sufficient data* – As we sought to use corpora of a sufficient density to offer input of representative diversity in terms of both vocabulary and sentence types, we included only those corpora which contained at least 10,000 words.
2) *Dyadic* – Because we wished to model production for each child, we selected only corpora which featured at least 1000 multiword child utterances, with a multiword child-to-adult utterance ratio of no less than 1:20.

These criteria were met by corpora for 160 individual children (Afrikaans: 2, Cantonese: 8, Catalan: 4, Croatian: 3, Danish: 2, Dutch: 12, Estonian: 3, Farsi: 2, French: 15, German: 22, Greek: 1, Hebrew: 6, Hungarian: 4, Indonesian: 8, Irish: 1, Italian: 8, Japanese: 10, Korean: 1, Mandarin: 7, Polish: 11, Portuguese: 2, Romanian: 1, Russian: 2, Sesotho: 3, Spanish: 11, Swedish: 5, Tamil: 1,

Welsh: 6). We recorded, for each child, the age range (mean age of 1;11 at the beginnings of the corpora, 3;10 at the ends), the number of months spanned by the corpus (mean: 23), the total number of words in the corpus (mean: 103,555), the number of child utterances (mean: 14,248), the number of multiword child utterances (mean: 7,552.7), the number of adult utterances (mean: 23,207), the multiword child-to-adult utterance ratio (mean: 0.49), and the mean words per child utterance (overall mean: 2.3). Since a method for automated morpheme segmentation was not available for all 28 languages, we do not include MLU calculations. For the full list of corpora (including references), see Appendix D.

The final set of 28 languages (including English, 29) differed typologically from one another in a number of important ways. Four dominant word orders were represented: SVO (18), VSO (2), SOV (4), and no dominant order (5; Haspelmath et al. 2005). The languages varied widely in their morphological complexity, ranging from languages with no morphological case marking (e.g., Sesotho; Demuth, 1992) to languages with 10 or more cases (e.g., Estonian; Haspelmath et al., 2005). Table 3 shows the family, genus, dominant word order, and number of cases for each of the 28 languages, in addition to English.

Table 3

Typological Properties of the 29 Languages

Language	**Family**	**Genus**	**Word Order**	**# Cases**
Irish	Indo-European	*Celtic*	VSO	2
Welsh	Indo-European	*Celtic*	VSO	0
English	Indo-European	*Germanic*	SVO	2
German	Indo-European	*Germanic*	N.D.	4
Afrikaans	Indo-European	*Germanic*	N.D.	0
Dutch	Indo-European	*Germanic*	N.D.	0
Danish	Indo-European	*Germanic*	SVO	2
Swedish	Indo-European	*Germanic*	SVO	2
Greek	Indo-European	*Greek*	N.D.	3

Farsi	Indo-European	*Iranian*	SOV	2
Romanian	Indo-European	*Romance*	SVO	2
Portuguese	Indo-European	*Romance*	SVO	0
Catalan	Indo-European	*Romance*	SVO	0
French	Indo-European	*Romance*	SVO	0
Spanish	Indo-European	*Romance*	SVO	0
Italian	Indo-European	*Romance*	SVO	0
Croatian	Indo-European	*Slavic*	SVO	5
Russian	Indo-European	*Slavic*	SVO	7
Polish	Indo-European	*Slavic*	SVO	7
Estonian	Uralic	*Finnic*	SVO	10+
Hungarian	Uralic	*Ugric*	N.D.	10+
Sesotho	Niger-Congo	*Bantoid*	SVO	0*
Hebrew	Afro-Asiatic	*Semitic*	SVO	0
Tamil	Dravidian	*S. Dravidian*	SOV	7 or 8**
Indonesian	Austronesian	*Malayic*	SVO	0
Cantonese	Sino-Tibetan	*Chinese*	SVO	0
Mandarin	Sino-Tibetan	*Chinese*	SVO	0
Korean	Korean	*Korean*	SOV	7
Japanese	Japanese	*Japanese*	SOV	9

Note: Information from Haspelmath et al. (2005), except where noted otherwise

**Demuth, 1992*

***Schiffman, 1999*

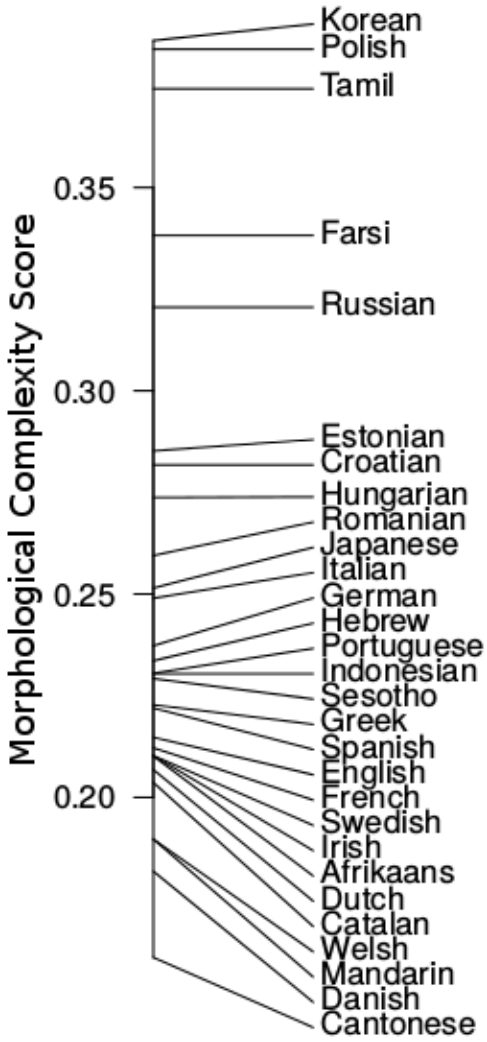
We sought to gauge the morphological complexity of the languages quantitatively, thereby placing them on an analytic/synthetic spectrum (Greenberg, 1960). Analytic languages such as Mandarin or English have a low morpheme-to-word ratio, whereas synthetic languages like Polish or Hungarian have a high morpheme-to-word ratio. We therefore carried out an analysis of the type/token ratio for each language (following Chang et al., 2008). This allowed us to approximate the morpheme-to-word ratio differences between languages without the aid of an automated method for morpheme segmentation: Morphological richness mirrors type/token ratio in the sense that morphologically complex languages yield a greater number of unique morpheme combinations, and thus a higher

number of unique word types, relative to the number of tokens, whereas analytic languages rely on a smaller number of unique morpheme combinations.

Thus, type/token ratio was used to compute a *Morphological Complexity* score for each language. For the type/token ratio calculation, we used only the adult utterances in the included corpora. Because type/token ratios are highly sensitive to the size of speech samples, we controlled for the lengths of individual corpora by calculating the mean type/token ratio per 2,000 words across all corpora representing a given language. The results of these calculations are depicted on an analytic/synthetic spectrum in Figure 9, and demonstrate the wide variation of the 29 languages, including English, in terms of morphological complexity. While some languages had relatively low *Morphological Complexity* scores (e.g., Cantonese), others had much higher scores (e.g., Tamil), and

others had scores falling between the two (e.g., Sesotho).[9]

Fig. 9: Morphological Complexity scores for each of the 29 languages.



[9] While type/token ratios allow us to gain a rough estimate of morphological complexity across all languages in the CHILDES database, the method is nonetheless highly sensitive to the properties of individual corpora. Note, for instance, that Catalan and Spanish diverge based on type/token ratio despite close similarities between the two languages. There is less data available for Catalan, and the four corpora which met our selection criteria have a lower mean target child age than the Spanish corpora – such factors will impact type/token ratio independently of the morphological properties of the language. Thus, these scores are meant merely to provide a very rough estimate of morphological richness.

Simulation 3: Modeling Child Comprehension of French and German

Shallow parsers (providing a gold standard for comprehension performance) were only available to us for two of the additional languages: French and German. TreeTagger (Schmid, 1995) was used to evaluate comprehension-related performance (through shallow parsing) for both languages. In this section, we report shallow parsing performance for French and German CBL simulations. Fifteen French and 22 German child corpora in the CHILDES database met our selection criteria and were used to simulate aspects of comprehension and production in exactly the same model architecture as used in the English simulations (as was also the case for the baseline models).

French: Comprehension performance. Across all 15 French single-child corpora, CBL achieved a mean F-score of 71.6, while the PARSER model reached a mean F-score of 64.4. The Trigram model attained a mean F-score of 59.0. Comprehension performance for each model is shown in Figure 10. As with the English simulation, the model not only outperformed its baselines, but yielded a tighter, more uniform distribution of scores.

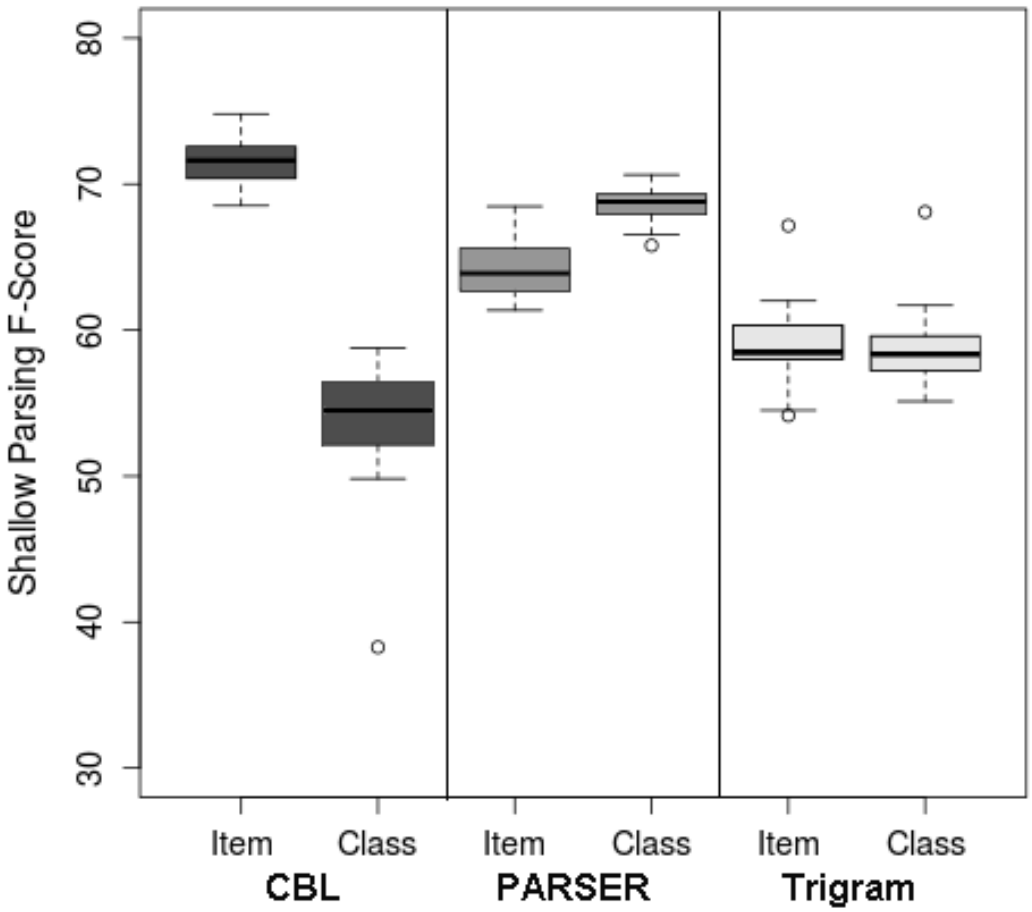


Fig. 10: Boxplots depicting shallow parsing F-score (%) for the CBL model and its baselines for the French simulations. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

The F-scores for the model and its baselines were logit-transformed and submitted to a repeated-measures ANOVA including the factor *Model* (2: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [$F(2,26) = 214.6$, $p < 0.0001$], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [$t(14)=14.33$, $p<0.0001$] and Trigram [$t(14)=15.72$, $p<0.0001$] models, as well as for PARSER compared to the

Trigram model [$t(14)=11.99$, $p=0.0001$].

As with the English simulations, we followed up this analysis by examining accuracy separately. Across the 15 child corpora, CBL attained a mean accuracy rate of 72.0%, while the PARSER model attained a mean accuracy rate of 61.8%. The Trigram model attained a mean accuracy rate of 57.0%. For completeness, CBL attained a mean score of 70.8%, while the PARSER model attained a mean completeness rate of 73.5%. The Trigram model attained a mean completeness rate of 66.1%. Detailed analysis of accuracy and completeness scores are provided in Appendix A.

Therefore, similar to our English simulations, the best combination of accuracy and completeness (as measured by the F-score), as well as the best accuracy specifically, was achieved by CBL for the French child corpora.

German: Comprehension performance. Across all 22 single-child corpora, CBL attained a mean F-score of 75.7, while the PARSER model attained a mean F-score of 73.4. The Trigram model reached a mean F-score of 67.4. Though CBL once more attained the highest scores, its performance advantage over the PARSER model was markedly smaller than in the English and French simulations. The distributions of scores for the model and its baselines are shown in Figure 11.

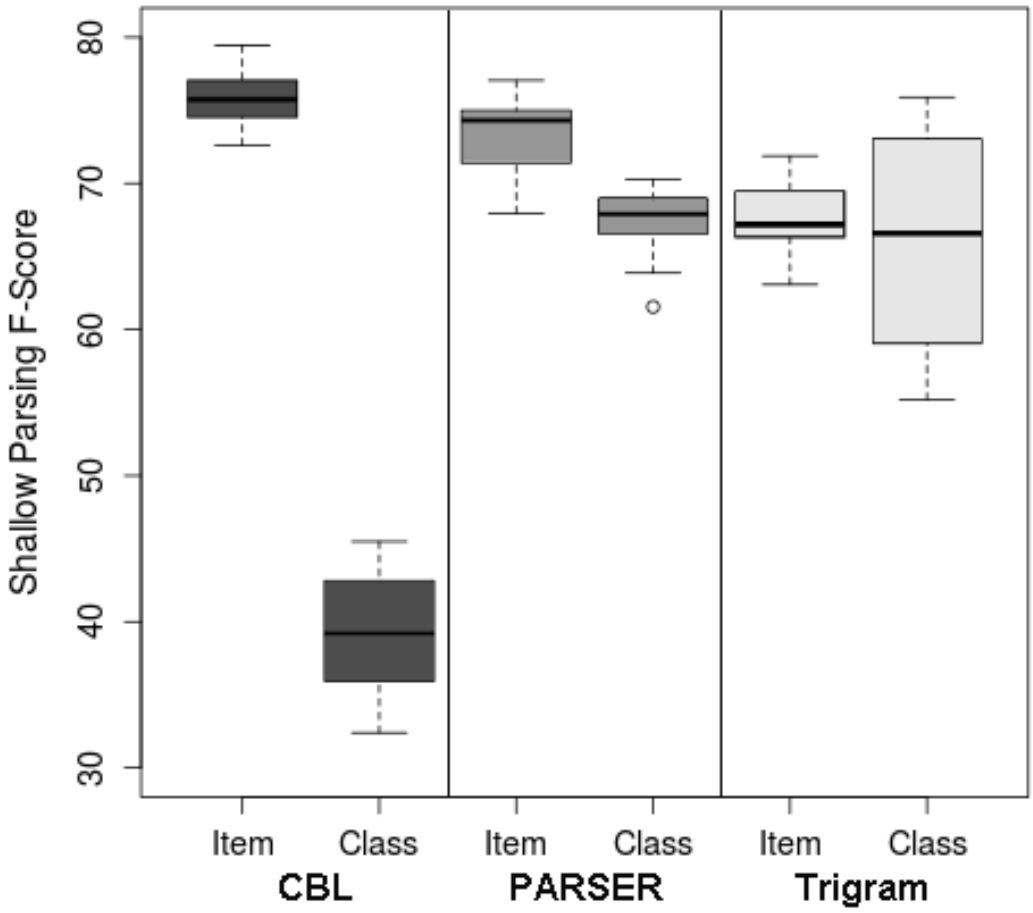


Fig. 11: Boxplots depicting shallow parsing F-score (%) for the CBL model and its baselines for the German simulations. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

The F-scores for the model and its baselines were once more logit-transformed and submitted to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [$F(2,40) = 69.43$, $p < 0.0001$],

with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [$t(21)=2.78$, $p<0.05$] and Trigram [$t(21)=17.67$, $p<0.001$] models, as well as for PARSER compared to the Trigram model [$t(21)=6.8$, $p=0.0001$].

As with the English and French simulations, we followed up this analysis by examining accuracy separately. Across the 22 child corpora, CBL attained a mean accuracy rate of 78.0%, while PARSER attained a mean accuracy rate of 69.4%. The Trigram model attained an accuracy of 70.5%. For completeness, CBL attained a mean score of 72.2%, while PARSER attained a mean completeness of 83.5%. The Trigram model attained a completeness of 62.9%. Accuracy and completeness scores are analyzed in Appendix A.

Thus, as with our English and French simulations, the best accuracy, as well as the best combination of accuracy and completeness (as measured by the F-score), was achieved by CBL across the German child corpora. PARSER tended to segment more heavily than with English and French, leading to a drop in accuracy, relative to baselines, and a boost in completeness.

Comparing CBL's French and German shallow parsing performance to English performance: The CBL model's performance was highly similar across English, French, and German. Beyond outperforming baseline models on all three languages (both in terms of Accuracy and in terms of the overall F-score), the model yielded mean scores which were highly similar across languages: mean F-scores of 75.4 (English), 71.6 (French), and 75.7 (German) were achieved, alongside mean Accuracy rates of 76.4 (English), 72 (French), and 78 (German) and mean completeness scores of 73.8 (English), 70.8 (French), and 72.2 (German).

Thus, the model's ability to group related words together in an incremental, on-line fashion was remarkably stable across the three languages, despite important differences along a number of dimensions such as morphological complexity (French and German are morphologically richer than English) and word order (while English and French have an SVO word order, German has no dominant

word order). This result offers important cross-linguistic support not only for the importance of multiword sequences, but for memory-based (as opposed to purely predictive) on-line learning as well as the plausibility of shallow linguistic processing based on local information.

Class-based simulations. We sought to test our item-based approach cross-linguistically by once more creating form class corpora from the single-child corpora used in the French and German simulations (using the same corpus preparation procedure described for English). We then tested the comprehension performance of the model and its baselines when learning from class-based statistics.

Figures 10 and 11 also compare class- and item-based F-scores for comprehension across the French and German corpora. For CBL, performance was considerably worse when working with class-based statistics, with a sharp decrease in the mean F-score for both French (from 71.6 to 53.5) and German (from 75.7 to 39.1). For PARSER, there was a similar decrease in performance for German (from 73.4 to 67.5), but an increase in score for French, from a mean F-score of 64.4 to 68.6. In the case of CBL, the lower comprehension performance when working with class statistics was driven by a drastic drop in completeness scores (French: from 70.8 to 29.9; German: from 72.2 to 21.4), owing partly to the use of the chunkatory in phrase segmentation; the relatively small number of possible class combinations in a sequence lead to the automatic grouping of items together (based on chunkatory searches) with increasing frequency throughout the models' pass through a corpus. As more combinations exceeded the average TP threshold, the models placed progressively fewer phrase boundaries. For CBL there were less drastic changes in accuracy (French: from 72 to 75.4; German: from 78 to 68.9).

PARSER actually increased French F-scores under class-based information, owing to an increase in accuracy (from 61.8 to 71.9), though with a drop in completeness (from 73.5 to 60.5), while German scores decreased due to a drop in completeness (from 83.5 to 56.9), while accuracy scores increased (from 69.4 to 74.8).

The Trigram baselines showed less drastic changes in performance: the mean F-score rose for French (from 59 to 66.2) while decreasing slightly for German (67.4 to 66). Though accuracy scores increased for sharply French (from 57 to 73.9), completeness dropped (from 66.1 to 51.5). German accuracy increased (70.5 to 76.9) while completeness also dropped (from 62.9 to 53.1).

We evaluated the effects of learning from French class-based information using a two-way ANOVA with factors *Statistic Type* (2: item- vs. class-based) and *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor, over logit-transformed F-scores. This yielded main effects of *Statistic Type* [$F(1,14) = 6.09$, $p<0.05$], confirming stronger performance for item-based models, and of *Model* [$F(2,28)=19.91$, $p<0.0001$], with an interaction between *Statistic Type* and *Model* [$F(2,28)=348.4$, $p<0.0001$], due to a more drastic drop in performance for the CBL model relative to the baselines when working with classes.

We also evaluated the effects of learning from German class-based information using a two-way ANOVA with factors *Statistic Type* (2: item- vs. class-based) and *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor, over logit-transformed F-scores. This yielded main effects of *Statistic Type* [$F(1,21) = 243.4$, $p<0.0001$], confirming stronger performance for item-based models, and of *Model* [$F(2,42)=324$, $p<0.0001$], with an interaction between *Statistic Type* and *Model* [$F(2,42)=301.3$, $p<0.0001$], due to a more drastic drop in performance for the CBL model relative to the baselines when working with classes.

As was the case with the English simulations, a reliance on word classes did not improve the performance of the model; instead, the use of classes leads to a decrease in performance. Though performance did increase for the baseline models when using French class-based information, CBL still yielded the strongest performance out of all simulations in its original item-based form. This result reaffirms the item-based approach, as well as the broader notion that initial language learning can take place without abstract syntactic categories. This also resonates with our previous simulation of child

artificial grammar learning (McCauley & Christiansen, 2011), casting further doubt on to claims that children and adults can use transitional probabilities to segment phrases by calculating statistics over word classes (e.g., Saffran, 2002; Thompson & Newport, 2007) rather than concrete items.

Chunk inventory characteristics across all 29 languages: While shallow parsing performance could only be automatically evaluated for English, French, and German, we nevertheless sought to explore differences in the development of the chunk inventory across the full set of languages. For each simulation, we separated the input corpus into 5 bins of equal size and plotted, separately for each bin, the percentage of actively used chunk types (weighted by frequency) consisting of multiple words (as in Figure 5). The outcome is depicted in Figure 12, organized based on the Morphological Complexity scores described above.

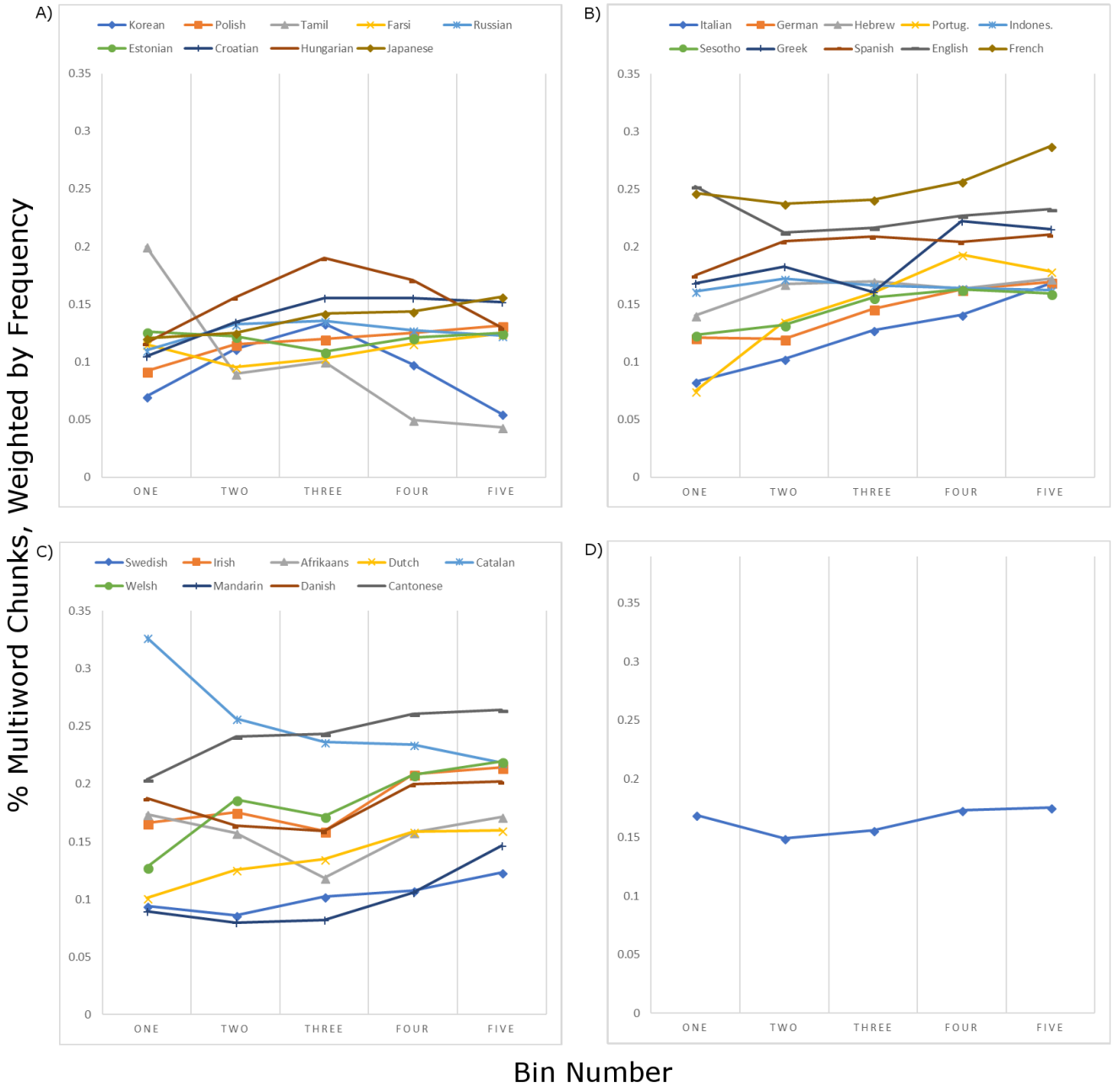


Fig. 12: Percentage of multiword chunk types weighted by frequency of use for: a) languages with high Morphological Complexity scores; b) languages with intermediate Morphological Complexity scores; c) languages with low Morphological Complexity scores; d) the overall mean across all 29 languages.[10]

As can be seen, there is considerable variation across languages. While the overall mean across bins

[10] Romanian was excluded from Figure 12 as it constituted an extreme outlier, with mean percentages of 83%, 11%, 19%, 38%, and 38% for each of the five successive bins.

roughly follows the U-shaped pattern observed for English, as do many of the individual languages, some of the languages (including two of the most morphologically rich: Korean and Hungarian) exhibit an inverted U-shape. The most consistent trend overall is that multiword chunk prominence increases over time, although some languages with extremely small corpora (Tamil, Catalan) exhibit the opposite pattern. Variation in corpus size is ultimately a limiting factor, and future work focusing on specific languages will be needed to determine the extent to which these differences stem from typological factors versus cultural differences in usage patterns versus individual differences across specific children and caretakers.

Additionally, we examined the relationship between morphological richness of the input language and the mean size of chunks (in words) discovered by the model. Unsurprisingly, there was a clear relationship between the two, as depicted in Figure 13.

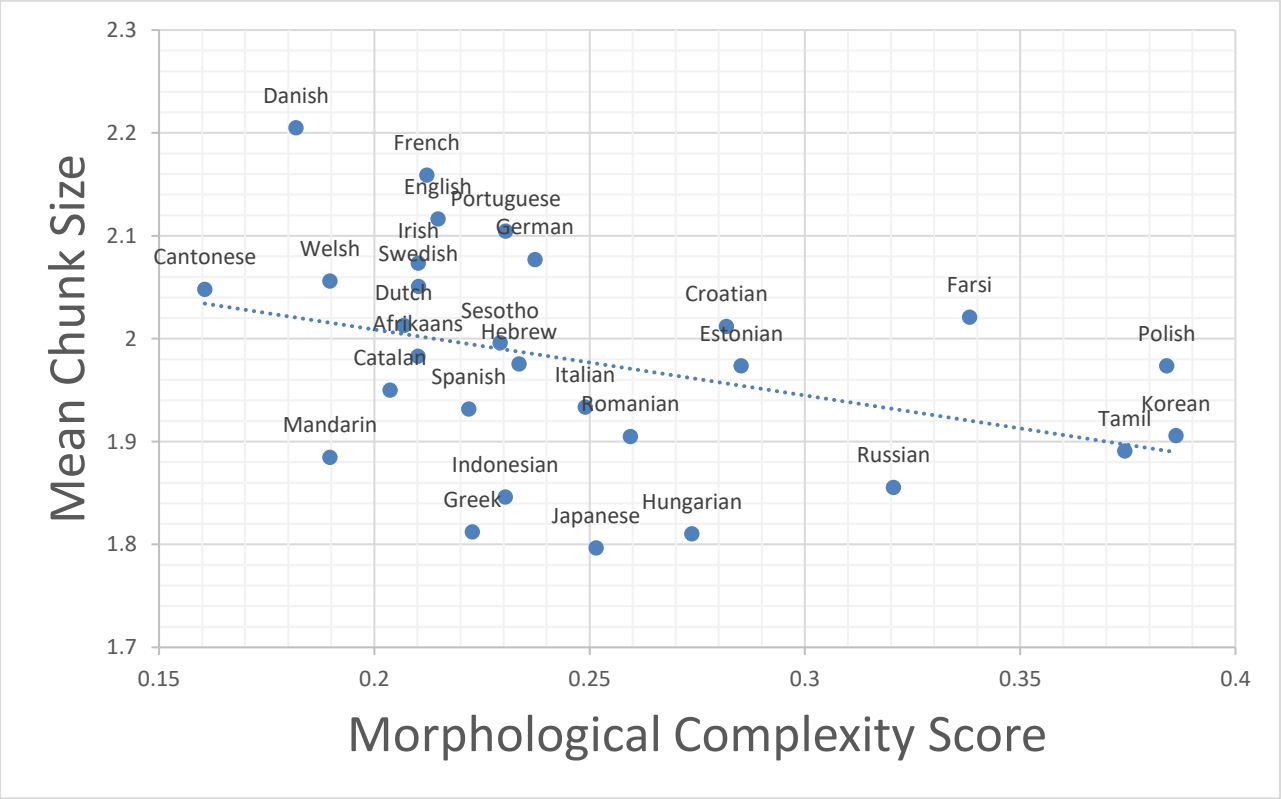


Fig. 13: Scatterplot depicting the relationship between mean chunk size and Morphological Complexity Scores of the

29 languages. Trendline derived from simple linear regression.

As can be seen, higher Morphological Complexity Scores predicted mean chunk size in the expected direction. However, this relationship was only marginally significant (β=-0.2, t=-2.05, p=0.051, R^2=0.134). Thus, there are clearly additional typological and usage differences across languages and individuals which go beyond morphological complexity (as estimated indirectly through type-token ratio) in determining the size of chunks discovered by the model.

Case study: Learning grammatical gender: Cross-linguistically, children master grammatical gender quite early in development (e.g., Slobin, 1986), and rarely make the sort of gender agreement errors often made by second language learners (e.g., Rogers, 1987; Holmes & de la Bâtie, 1999). Such findings resonate with the proposal that children treat article-noun pairs as single units (e.g., MacWhinney, 1978; Carroll, 1989), an idea which receives support from item-based patterns observed in children's use of articles (e.g., Mariscal, 2008; Pine & Lieven, 1997). More recently, Arnon and Ramscar (2012) used an artificial language learning paradigm to test the idea that learning article-noun pairings as chunks imparts an advantage in the learning of grammatical gender. They found that subjects receiving initial exposure to unsegmented article-noun sequences, which was only later followed by exposure to the noun labels in isolation, exhibited better mastery of grammatical gender in the artificial language at test than did those subjects who underwent the very same exposure phases in reverse order.

The findings of Arnon & Ramscar (2012), as well as children's item-based patterns in article usage and the apparent ease with which they master grammatical gender more generally, lead us to examine the model's ability to construct the right article-noun pairings during production. While CBL does not possess chunks arrived at via under-segmentation (the model initially recognizes articles and nouns as separate entities by virtue of the fact that the input corpora are in the form of words), the

model may nevertheless learn to chunk articles and nouns together, leading to an item-based mastery of grammatical gender. To explore the model's learning of grammatical gender and determine whether its production of article-noun pairs exhibits the early mastery demonstrated by children, we analyzed the model's productions from the simulation involving the largest German corpus from CHILDES (Leo). We found that the model's article-noun pairings were correct over 95% of the time. Out of those article-noun pairs produced by the target child in the corpus[11], the model correctly captured 11,703, pairing the wrong article with a noun in only 557 cases. When we considered only those 513 utterances which featured multiple articles (as in the sentence ***die*** *Katze jagte* ***den*** *Hund*), rather than two or more instances of the same article being paired with different nouns (as in ***die*** *Katze jagte* ***die*** *Maus*), we found that the model paired nouns with the wrong gender marker in only 14 cases (an error rate of 2.7%). Thus, consistent with the findings of Arnon & Ramscar (2012), the distributional learning of article-noun sequences as chunks leads the model to mirror both children's early mastery of grammatical gender as well as the item-based nature of children's early article usage.

Simulation 4: Modeling Child Production Performance across a Typologically Diverse Set of 29 Languages

[11] In theory, this could also include incorrect article-noun pairings produced by the target child of the corpus, but previous work (e.g., Rogers, 1987; Holmes & de la Bâtie, 1999) suggests children rarely make errors with grammatical gender.

We conducted production simulations for each corpus from the additional languages, using the same model architecture and baseline models as used in each of the previous simulations. In the overall analysis of Sentence Production Performance, we include the scores for the English corpora for a total of 204 individual child simulations. CBL achieved a mean sentence production accuracy of 55.3%, while the Trigram model achieved a mean sentence production accuracy of 46%. The results for each language are depicted in Figure 14.

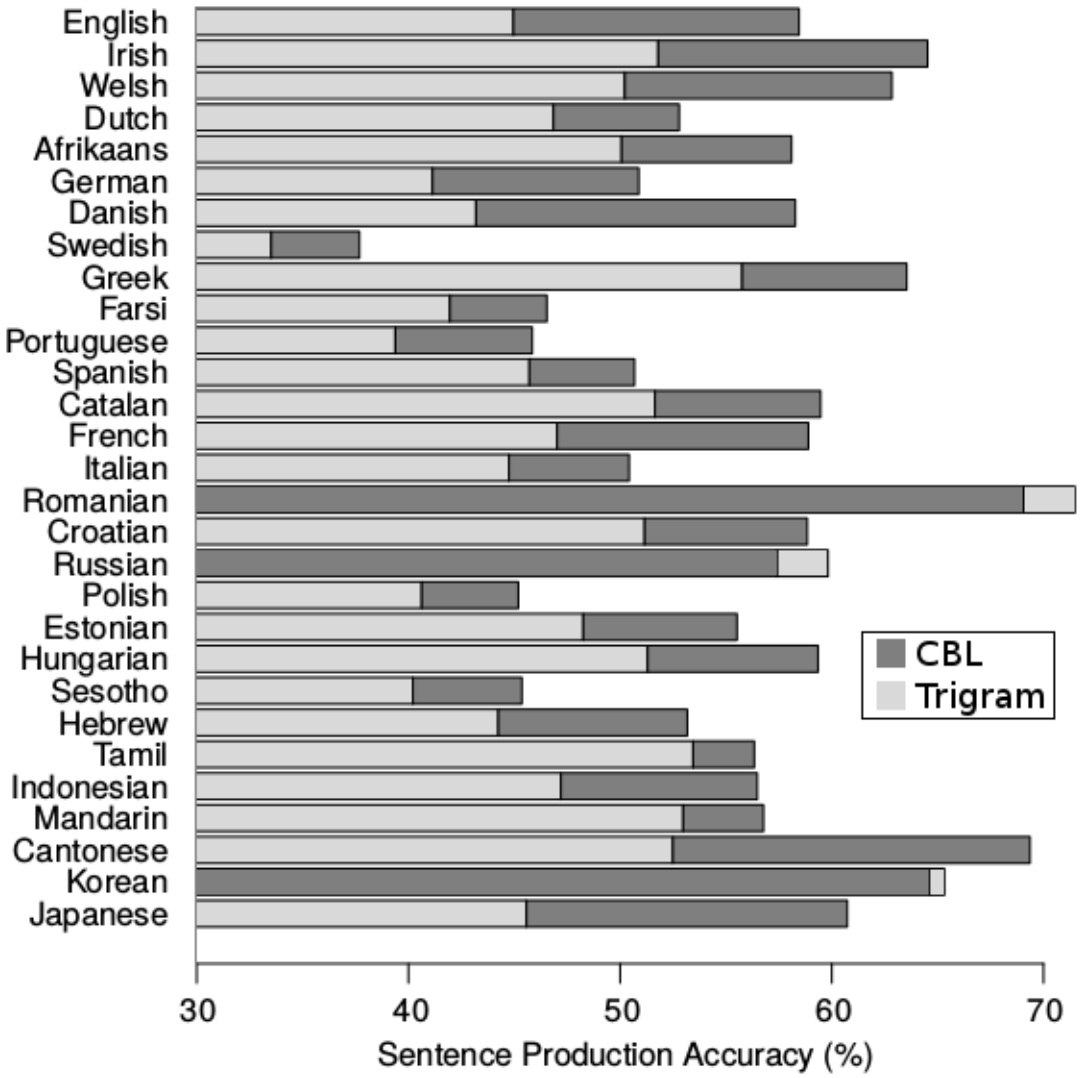


Fig. 14: Mean Sentence Production Accuracy scores for the CBL model and its trigram baseline across all 29 languages, including English (shown at top). Bars are non-cumulative (e.g., the Japanese CBL score was just over

60%, while the Trigram score was near 45%).

As with the English-only production simulations, we submitted the production scores to a repeated-measures ANOVA including, once more, the factor *Model* (2: CBL vs. Trigram) with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [$F(1,203) = 575.2$, $p < 0.0001$], indicating better performance for CBL.

When all utterances across the simulations were considered together, CBL was able to produce the majority of those utterances. The same pattern held for all but five of the individual languages, with CBL failing reach the 50% mark for Sesotho, Polish, Farsi, Portuguese, and Swedish.

As can be seen, CBL outperformed its baseline for 26 of the 29 languages; the exceptions were Russian, Romanian, and Korean for which the Trigram scored highest. It should be noted that for two of the exceptions (Romanian and Korean), there was only one child corpus; for Russian, there were only two available corpora. Moreover, all three of these languages fall towards the extreme synthetic end of the analytic/synthetic spectrum estimated by our morphological analyses. Below, we explore the notion that the CBL model performed worse as a function of morphological complexity.

Effects of morphological complexity and word order: To assess the effect of morphological complexity on the model's performance, we fit a linear regression model to the Sentence Production Accuracy scores across the 204 simulations using the morphological complexity measure calculated for each language previously. This yielded a significant negative value for morphological complexity [β=-0.17, $t(202)$=-4.35, $p<0.0001$], indicating that the model's sentence production tended to be less accurate when learning morphologically rich languages, although the amount of variance explained by the linear model was moderate to low (Adjusted R-squared: 0.08). Figure 15 depicts the Sentence Production Accuracy for each simulation according to the morphological complexity score of the 29

languages.

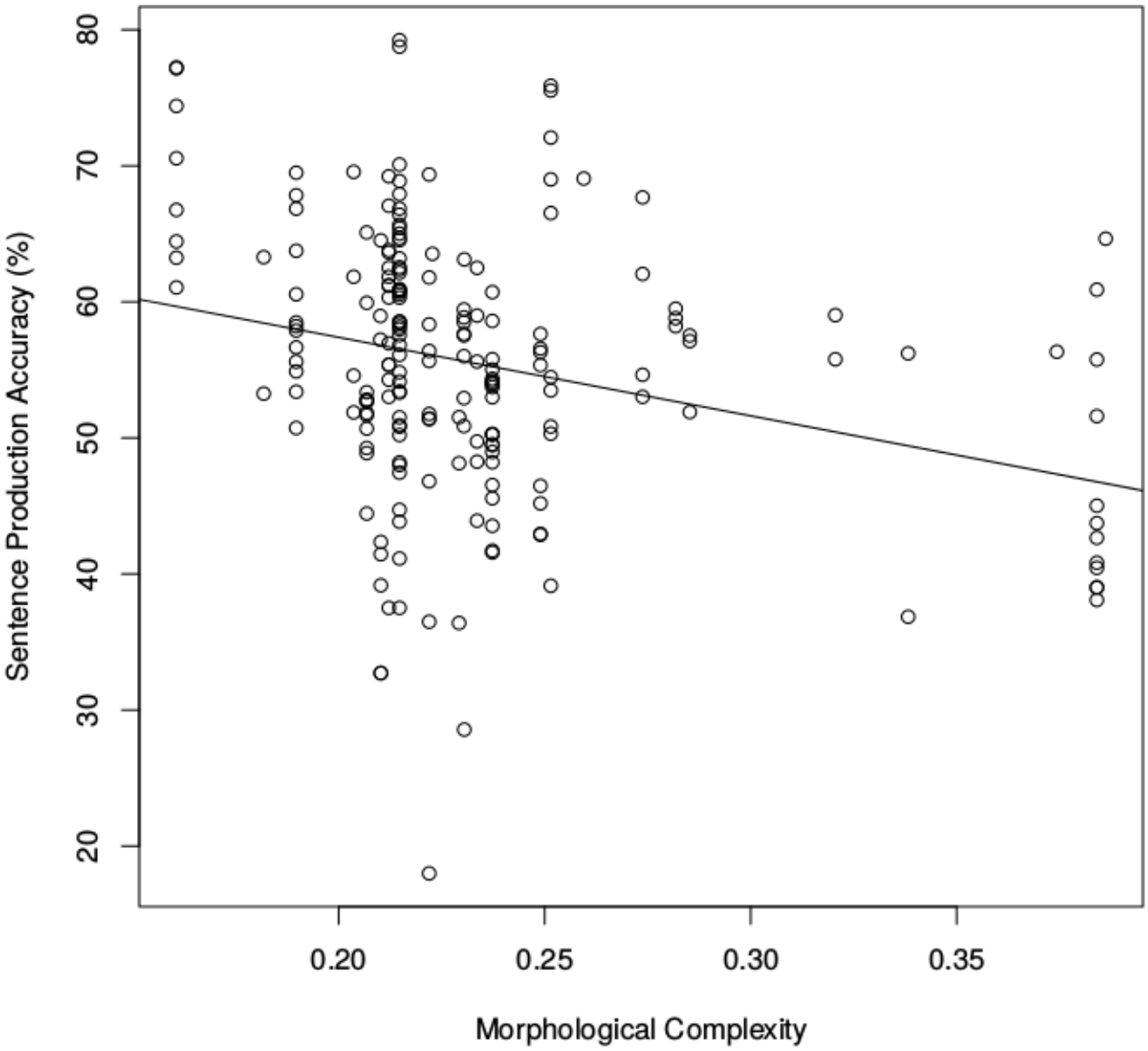


Fig. 15: Scatterplot depicting Sentence Production Accuracy (%) for each simulation by the Morphological Complexity score for the corresponding language.

Because the three languages on which the word-based Trigram model outperformed CBL were on the extreme synthetic end of the analytic/synthetic spectrum, we tested whether the model's sentence production accuracy advantage over the Trigram baseline decreased as a function of morphological richness. A linear model with *Morphological Complexity* as a predictor of the difference between CBL and Trigram scores for each simulation confirmed this to be the case [β=-0.4, $t(201)$=-5.6, $p<0.0001$], with the overall model explaining a significant amount of variance in the difference scores (Adjusted R-squared: 0.13).

That the advantage of a chunk-based model such as CBL would decrease as a function of morphological complexity is perhaps unsurprising. However, segmenting corpora into their component morphemes may better accommodate a chunk-based model while helping to deal with data sparseness tied to high type/token ratios. Since an automated method for morpheme segmentation was not available for many of the languages, we were unable to test this intuition cross-linguistically. However, the CHILDES segmentation system for one of the more synthetic languages, Japanese, treats verb morphemes as separate words (Chang et al., 2008; Miyata, 2000; Miyata & Naka, 1998). Interestingly, model performance, on average, is far stronger for Japanese corpora than for languages of comparable morphological complexity, especially relative to the Trigram baseline. Additionally, the greatest difference in scores between the model and its baselines was seen for Japanese. Future work will focus on comparing model performance on synthetic languages with morphologically segmented vs. standard corpora.

Despite the effect of morphological complexity on model production performance, there was no effect of word order (mean scores: SOV, 58.7; SVO, 55.5; VSO, 63; no dominant order: 52.9). We used a linear model to test for a potential effect of word order on model performance, while controlling for morphological richness. As the effect of *Morphological Complexity* on model performance was

significant, we included it as a predictor alongside *Word Order* in the model. However, no significant effect of word order emerged. Figure 16 depicts the mean Sentence Production Accuracy score across each of the four word orders.

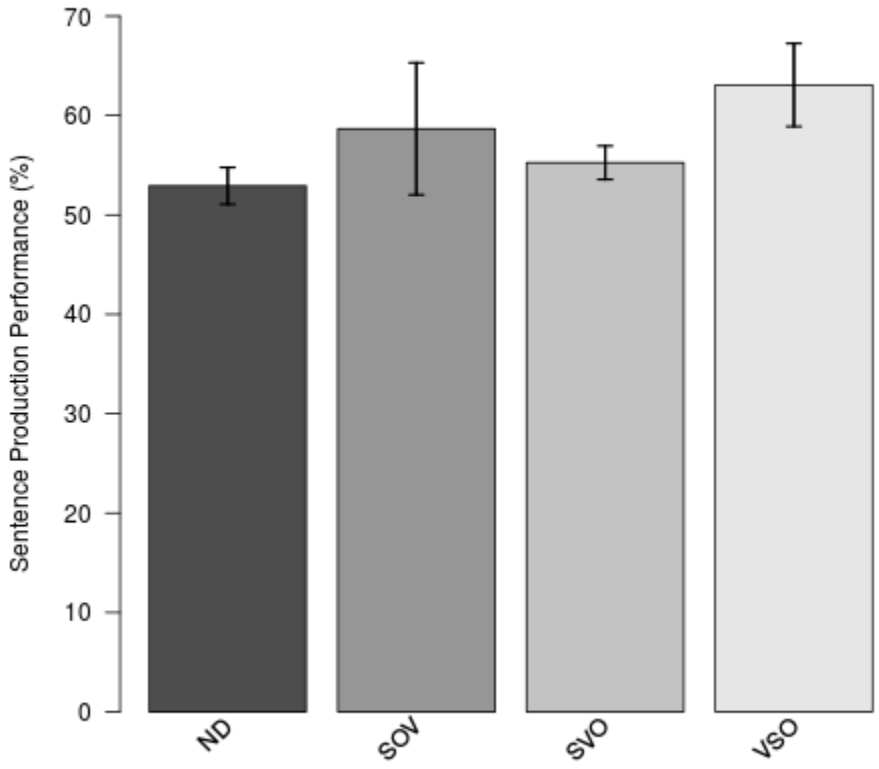


Fig. 16: Barplot depicting the mean Sentence Production Accuracy (%) for each of the four word orders represented across the 29 languages. Error bars depict standard errors. While there were only two corpora tested with VSO word order (precluding a statistical test), visual inspection of the error bars indicates that the model's performance was highly similar across the four word orders represented.

Production summary: The model outperformed its baseline for 26 of the 29 languages, correctly producing the majority of child utterances for 24 languages. The usefulness of BTP-based chunking across so wide an array of languages is somewhat surprising, given previous work

demonstrating that the usefulness of forward vs. backward probability calculations in word segmentation is highly sensitive to cross-linguistic differences in distributional features (e.g., heavy suffixing in case of Hungarian, phrase-initial function words in the case of Italian; Gervain & Guevara Erra, 2012). While our model was somewhat sensitive to differences in morphological complexity, tending to perform slightly better on morphologically simple languages, it did not appear to be sensitive to differences in dominant word order. The fact that multi-word units were useful even for the learning of morphologically rich languages is of particular interest, considering the difficulties inherent in working with morphologically rich languages in the field of computational linguistics (cf. Tsarfaty, Seddah, Kübler, & Nivre, 2012). Taken together with previous findings of item-based patterns in children's learning of morphologically rich languages (e.g., MacWhinney, 1978), this result is quite encouraging in the context of future cross-linguistic item-based modeling work.

These results offer substantial cross-linguistic support for CBL, and, more broadly, for the view that simple learning mechanisms underlie a large part of early linguistic behavior. The outcome of our simulations strengthens previous psycholinguistic evidence for chunk-based learning, which has been gained primarily from English speakers (e.g., Bannard & Matthews, 2008), suggesting that multiword sequences play an important role cross-linguistically, in analytic and synthetic languages alike.

General Discussion

We have shown that the CBL model can approximate key aspects of children's comprehension and production of language by learning in a purely incremental fashion through on-line processing. The model gradually builds up an inventory of chunks consisting of one or more words, which unites aspects of comprehension and production within a single framework. On the comprehension side, the

model chunks incoming words together to incrementally build an item-based “shallow parse” of each utterance as it is encountered. Chunks discovered in this fashion are used to make predictions about upcoming words in subsequent input, facilitating the model's on-line processing. When the model encounters an utterance produced by the target child of the input corpus, it attempts to generate an identical utterance using the same chunks and statistics used in shallow parsing. Importantly, production is modeled as an incremental, chunk-to-chunk process rather than one of whole-sentence optimization (as would be the case by, e.g., choosing among candidate sentences based on whole-string probabilities).

The model achieves strong performance across English single-child corpora from the CHILDES database, approximating the performance of a shallow parser with high accuracy and completeness. In line with expectations derived from usage-based theories, item-based information is shown to be more useful than statistics or chunks computed over form classes. The model is also able to fit much of children's early linguistic behavior, correctly generating the majority of the child utterances encountered across the English corpora. The model exhibits similar shallow parsing performance for 15 French and 22 German corpora alongside similar sentence production performance for nearly 200 additional child corpora drawn from a typologically diverse set of 28 languages (also from CHILDES). In each case, the model outperforms baseline models. In addition to its strong cross-linguistic performance in approximating aspects of comprehension and production, the model provides close quantitative fits to children's production of complex sentences featuring six different relative clause types (Diessel & Tomasello, 2005).

Together, these findings suggest that a fair amount of children's early language use may be supported by incremental, on-line learning of item-based information using simple distributional cues. By the same token, the model also serves to highlight the limits of what can be achieved through

distributional learning alone. In what follows, we discuss the limitations of the model and directions for future modeling work addressing them. We then place the model in the larger context of usage-based computational approaches to acquisition. Finally, we highlight insights drawn from the model to be explored in future psycholinguistic work.

Limitations of the Model

As an initial step towards a comprehensive computational account of language learning, CBL is not without limitations. Perhaps most immediately obvious is that the model learns from segmented input. It does not confront one of the early challenges facing language learners: that of segmenting an uninterrupted speech stream. The problem of segmenting the speech stream (traditionally thought of as word segmentation) and discovering useful multiword sequences are likely to impact one another: Children do not learn about phonology, words, multiword sequences, and meanings in discrete, separable stages, but instead learn about multiple aspects of linguistic structure simultaneously and their interaction with each other (see also Christiansen & Chater, 2016b). Indeed, many of children's earliest, unanalyzed chunks are likely to stem from under-segmentation “errors,” which may offer insights into a number of phenomena tied to early language learning (Arnon, 2009; Arnon & Christiansen, 2017). Future work will focus on using the model to learn from unsegmented corpora in ways that maintain a fluid rather than rigid relationship between individual words, unanalyzed chunks, and chunks which the model is capable of breaking down into its component words.

A further limitation stems from what may also arguably be one of the model's greatest strengths: reliance on a single source of distributional information. CBL was designed to be as simple as possible, in order to demonstrate that a model can approximate aspects of comprehension and production through incremental, on-line processing based on simple statistics. Though the model, which relies

upon BTPs, is evaluated against a baseline which uses FTPs in Appendix C, it is clear from the modeling results that both information sources are potentially useful. Infants, children, and adults have been shown to be sensitive to TPs calculated in both directions (e.g., French et al., 2011; Pelucchi et al., 2009; Perruchet & Desaulty, 2008). Future work should be based on a principled, parameter-free method for seamlessly integrating TP calculations in both directions (in addition to other potentially useful distributional and non-distributional cues).

A further limitation is demonstrated by work with adult subjects, which suggests that there is no frequency "threshold" beyond which a multiword sequence is stored as a chunk, but rather that the extent to which sequences cohere as multiword units is graded in nature (cf. Caldwell-Harris et al., 2012). While CBL does not make use of raw whole-sequence frequency information in chunk discovery, it does rely on the use of a running average BTP as a threshold. Future work might benefit from considering the graded nature of "chunk" status for multiword units, while also seeking to make predictions about part/whole interactions (reflecting findings that stored multiword sequences both prime and are primed by their component words; e.g., Sprenger et al., 2006).

CBL's lack of "abstraction" over chunks: The current lack of abstraction in the model leads to a number of limitations, such as its inability to produce utterances "from scratch:" The randomly ordered bag-of-words which the model attempts to sequence during production is always populated by words from one of the target child's actual utterances. This means that the model cannot be used to capture children's systematic errors on a case-by-case basis (the model can only commit errors which are made by the target child). Previous models capable of producing novel utterances have successfully captured such developmental trends, such as optional infinitive errors (e.g., MOSAIC; Freudenthal et al., 2006, 2007).

A number of previous models have served to demonstrate that considerable linguistic

productivity can emerge from abstracting over multiword sequences (e.g., Solan et al., 2005). Thus, ongoing work with CBL seeks to derive partially-abstract, item-based schemas ("I want more X," where X is restricted to a class of words the model has learned to group together) while still adhering to the psychological principles of incremental, on-line processing without automatic storage of global utterance properties. The learning of increasingly abstract units may be essential in boosting model performance on the shallow parsing and production tasks to the limits of what is possible using distributional information alone.

Moving beyond the limitations of a purely distributional approach: Each of the limitations discussed immediately above can be addressed within a purely distributional framework. From a multiple-cue integration perspective (e.g., Bates & MacWhinney, 1989; Monaghan & Christiansen, 2008), distributional information is only one of a number of crucial factors in the language learner's input. The present study goes some distance towards demonstrating how far a single source of distributional information can take the learner while also underscoring the need for moving beyond a purely distributional framework; it attains strong performance in capturing specific aspects of language learning and use, but cannot hope to offer a more complete account of comprehension and production.

Thus, a limitation which must be addressed lies in the model's lack of semantic information; the model never learns "meanings" corresponding to the chunks it discovers, and is never called to interpret the meanings of utterances. Moreover, the psychological motivation for the model is partially driven by the perspective that semantic/conceptual information—such as that tied to event schemas, situational settings, and background world knowledge—is a key factor in generalizing to unbounded productivity of the sort exhibited by mature language users, superseding the importance of abstract "syntactic" knowledge, such as that of form classes (as discussed in the Introduction). The project of expanding the model to incorporate such information in idealized forms therefore represents a key

challenge.

One of the most tractable aspects of meaning for an incremental, on-line model such as CBL to learn and use lies in the assignment of semantic roles (often referred to as thematic roles), such as AGENT, ACTION, and PATIENT, and their use in the development of verb-argument structure constructions. Support for the psychological reality of semantic roles comes from empirical work on adult sentence comprehension (e.g., Altman & Kamide, 1999; Carlson & Tanenhaus, 1988). A number of the earliest computational models of language to incorporate meaning were focused on learning to assign semantic roles to sentence constituents within a connectionist framework (McClelland & Kawamoto, 1986; St. John & McClelland, 1990), while more recent connectionist modeling has extended the use of semantic roles to networks that use them to acquire verb-argument structure and make contact with a range of psycholinguistic data related to meaning (e.g., Chang et al., 2006). Such models are limited, of course, in that they are trained on non-naturalistic datasets rather than full corpora of child-directed speech, and in the case of the Chang et al. (2006) model, the problem facing the learner is simplified considerably by assuming the correct mapping between roles and lexical-semantic representations. Nevertheless, such approaches demonstrate the feasibility and psychological value of semantic role information in capturing meaning in comprehension and production (for an extensive review of models that deal with semantic roles and argument structure, see McCauley & Christiansen, 2014b).

However, it remains unclear whether young children possess coarse-grained, canonical semantic roles such as AGENT, ACTION, and PATIENT (cf. Shayan, 2008, for a review and empirical data), with some researchers going so far as to suggest that even adults represent thematic roles in a verb-specific manner (McRae, Ferretti, & Amyote, 1997). Thus, the extension of CBL and similar models should ideally involve the learning of semantic roles from the input, rather than use canonical

roles that are pre-determined and fixed. Alishahi and Stevenson (2010) achieve an initial step in this direction with a model that learns a probability distribution over featural semantic properties of arguments, which allows semantic roles and verb-argument structure to develop simultaneously. Featural input for nouns is derived from WordNet (Miller, 1990), which yields lists ranging from specific to general properties (e.g., *CAKE:* {baked goods, food, solid, substance, matter, entity}), while verbs involve hand-constructed primitives (e.g., *EAT:* {act, consume}). Throughout the course of exposure to the input corpus, the model gradually sees the transformation of item-based roles into more abstract representations which capture semantic properties of arguments across a range of verbs.

Thus, a promising initial step for CBL in moving towards a more comprehensive account of comprehension and production lies in the use of automatically generated featural input (utilizing existing resources such as FrameNet: Baker, Fillmore, & Lowe, 1998; VerbNet: Kipper-Schuler, 2005; and WordNet: Miller, 1990), which is then presented as input to the model alongside corresponding utterances in a child corpus. The key psychological underpinnings of the model can be maintained by ensuring that approximations of semantic roles, argument structures, etc., are learned through simple statistical and recognition-based processes that can be carried out incrementally. For instance, the recognition-based "prediction" mechanism currently featured in CBL could be slightly modified to accomplish something similar to the alignment and comparison technique of Solan et al. (2005) in an on-line fashion. The resulting item-based schemas could be further refined through learning featural information to arrive at partially-abstract constructions. The notion that an extended version of CBL could accomplish something on this level without resorting to probabilistic inference is bolstered by a recent model of reading (Baayen, Milin, Durdevic, & Hendrix, 2011) which is able to account for a range of psycholinguistic data through associative learning processes tying letters and letter trigrams to meaning representations derived from morphemes.

Relationship to other Usage-Based Modeling Approaches

Despite its current limitations, the CBL model may be viewed as a possible foundation for a more comprehensive computational account of language acquisition. While previous computational models within the usage-based tradition have boasted great success, CBL possesses a number of features that have been largely absent from language modeling, several of which represent desiderata for a fully comprehensive approach to acquisition.

Firstly, and perhaps most importantly, CBL takes usage-based theory to its natural conclusion in making no distinction between language learning and language use (Chater & Christiansen, in press); the model learns solely by attempting to comprehend and produce language. That is, the very processes by which input is interpreted and output is constructed are the same processes involved in learning; at no point does the model engage in a separate "grammar induction" process. This sets the present model apart from a number of extant usage-based models that have focused on grammar induction (e.g., Bannard et al., 2009) or conceived of learning and processing separately.

Also of great importance is that CBL learns incrementally, without batch learning of the sort used by most existing computational approaches (e.g., Bannard et al., 2009; Jones et al., 2004). While more sophisticated models of grammatical development have captured incremental learning (e.g., Bod, 2009; Kolodny et al., 2015), CBL is unique in offering an account of the on-line processes leading to linguistic knowledge over time; the model learns incrementally not only from utterance-to-utterance, but within individual utterances themselves as input is received, on a word-by-word basis. Thus, its design reflects the constraints imposed by the Now-or-Never bottleneck (Christiansen & Chater,

2016b).

Although a number of previous usage-based models have captured the generation of novel utterances (e.g., Jones et al., 2004; Solan et al., 2005), none have simultaneously sought to approximate aspects of comprehension in an explicit fashion. A further contribution of CBL is that it not only captures aspects of both comprehension and production, but also in unites them within a single framework. Pickering and Garrod (2007; 2013) argue that comprehension and production should not be seen as separate processes, a view compatible with usage-based approaches more generally (cf. Chater et al., 2016; McCauley & Christiansen, 2013). While connectionist models have utilized the same network of nodes to simulate comprehension and production (e.g., Chang et al., 2006; see also MacKay, 1982), ours is the first full-scale (taking full corpora as input) model to offer a unified framework.

Finally, CBL was designed to reflect psychologically parsimonious processing and knowledge representation. Outside the realm of word segmentation, the model is unique in its reliance solely on simple recognition-based processing and simple statistics of a sort that infants, children, and adults have been shown to be sensitive to (BTPs; French et al., 2012; Pelucchi et al., 2009; Perruchet & Desaulty, 2008). While a number of more complex computational approaches have made use of transitional probabilities (e.g., Kolodny et al., 2015; Solan et al., 2005), CBL relies solely on transitional probabilities computed in an incremental, on-line fashion, and is not supplemented by more complex processes. Furthermore, the model relies on local information; rather than automatically storing entire utterances, the model shallow parses and produces utterances in an incremental, chunk-to-chunk fashion rather than relying on whole-sentence representation or optimization.

Insights Derived from the Model

Beyond CBL's unique features, its ability to capture much of children's early linguistic behavior cross-linguistically, and its success in accounting for key psycholinguistic findings, the model leads to several insights which may be further explored through psycholinguistic research:

1) **Simple distributional cues are useful at every level of language learning.** The model was able to use a simple distributional cue previously shown to be useful in word segmentation (BTP; Pelucchi et al., 2009; Perruchet & Desaulty, 2008) in order to segment speech into useful multiword units, as well as to combine them to create sentences. Though based on this simple statistic, the model was able to make close contact with psycholinguistic results on children's production of complex sentence types (Diessel & Tomasello, 2005).

2) **Previous artificial grammar learning results may reflect item-based rather than class-based computations.** The decision to focus on learning through purely item-based statistics stands in contrast to several threads of argument within the statistical learning literature, which hold that learners discover phrase structure by computing statistics over form classes rather than individual words (e.g., Saffran, 2002; Thompson & Newport, 2007). As discussed above, we have found that the discovery of useful chunks of local information (which, being as our model was scored against a shallow parser, is analogous to phrase segmentation of the sort discussed by Thompson & Newport, 2007) was actually enhanced as a consequence of a reliance on item-based statistics, as was also often the case with the model's baselines. The model performed worse when exposed to class-based information – a pattern which was replicated in our simulation of Saffran's (2002) child artificial language learning experiment (McCauley and Christiansen, 2011).

We suggest, then, that artificial language learning results which have previously been taken to reflect learners' sensitivity to the phrase-like structure of the stimuli be reassessed to

determine whether item-based calculations might be sufficient to capture the learning of both children and adults. Previous modeling work on chunking has been shown to better account for segmentation performance in artificial language studies than more complex learning mechanisms (e.g., Perruchet et al., 2002; French et al., 2011). This general approach may be extended to full-blown language development.

3) **Most of the difficulty faced by the learner lies outside the distributional realm.** The difficulty of learning from distributional information may be compounded by the problem combining multiple probabilistic cues (which CBL, relying on a single distributional cue, does not attempt to capture). However, given the rapidity with which the model was able to learn how to identify useful chunks of local information, as well as to sequence those chunks to create new utterances, we suggest that the greatest difficulties children face in learning to process sentences may have less to do with distributional information or even "linguistic structure," but instead derive from conceptual/semantic dimensions of the problem, such as learning event schemas and scenarios to map chunks onto.

4) **Multiword sequences remain important throughout development.** The model ultimately relied more heavily on multiword units over time, as shown in various analyses of the model's chunk inventory. This leads us to suggest that instead of "starting big" by merely relying upon multiword units during the early stages of learning (e.g., Arnon, 2009), learners continue to rely on chunks throughout development; representations may actually become *more* chunk-like instead of less. At the same time, subtle shifts in the "degree of chunkedness" of children's representations, as suggested by the U-shaped development of the model's chunk inventory, may interact with U-shaped trajectories observed in seemingly disconnected areas, such as the learning of irregular forms.

5) **Learners rely on multiword units even in morphologically rich languages.** The model benefited from the use of chunks in learning analytic and synthetic languages alike. At the same time, chunk-to-chunk sequential information of the type learned by the model clearly matters less in synthetic languages, where there may be stronger pragmatic constraints on ordering. Moreover, as suggested by a corpus analysis of Turkish (Durrant, 2013)—an agglutinating language—chunking over sublexical elements, such as morphemes, might be important to the processing of morphologically rich languages. CBL currently lacks such information and this may partly explain the lower performance of the model when learning morphologically rich languages.

Conclusion

We have presented the foundations of a new approach to modeling language learning in the form of the CBL model, which provides a computational framework based on incremental, on-line learning from simple chunks and statistics. The model makes close contact with psycholinguistic evidence for both multiword unit storage and shallow, underspecified language processing; rather than attempting to induce a target grammar, the model learns chunks of local information which are used to simulate aspects of comprehension and production. CBL approximates the performance of a shallow parser by segmenting utterances into chunks of related words on-line, and simultaneously uses the same chunks to incrementally produce new utterances. The model's production abilities can account for a considerable part of children's early linguistic behavior. The model offers broad, cross-linguistic coverage and successfully accounts for key developmental psycholinguistic findings, in addition to making several predictions on which to base subsequent psycholinguistic work. But, perhaps most importantly, CBL demonstrates how language learning can modeled as language use.

Acknowledgements

We thank Pierre Perruchet for discussions relating to our implementation of the PARSER computational model. We are also grateful for the comments of six anonymous reviewers on previous versions of this paper. This work was supported in part by BSF grant number 2011107 awarded to MHC (and Inbal Arnon).

Preliminary versions of some of the results presented in this paper have been presented at the 33rd Annual Meeting of the Cognitive Science Society (2011), 53rd Annual Meeting of the Psychonomic Society (2012), the 12th International Cognitive Linguistics Conference (2013), the International Congress for the Study of Child Language (2014), the 8th International Conference on Construction Grammar (2014), and the Architectures and Mechanisms of Language Processing conference (2017).

REFERENCES

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*, 275–290.

Akhtar, N. (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of Child Language, 26,* 261–278.

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*, 789–834.

Alishahi, A., & Stevenson, S. (2010). Learning general properties of semantic roles from usage data: A computational model. *Language and Cognitive Processes, 25*, 50-93.

Ambridge, B., Rowland, C. F., & Pine, J. M. (2008). Is structure dependence an innate constraint? New experimental evidence from children's complex question production. *Cognitive Science, 32*, 222-255.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247-264.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30,* 191-238.

Arnon, I. (2009). Starting Big: The role of multi-word phrases in language learning and use. Unpublished doctoral dissertation. Stanford University, Palo Alto.

Arnon, I. & Christiansen, M.H. (2017). The role of multiword building blocks in explaining L1-L2 differences. *Topics in Cognitive Science, 9,* 621–636.

Arnon, I., & Clark, E. (2011). Why brush your teeth is better than teeth: Children's word production is facilitated by familiar frames. *Language Learning and Development, 7*, 107-129.

Arnon, I., & Cohen Priva, U. C. (2013). More than words: The effect of multi-word frequency and

constituency on phonetic duration. *Language and Speech*, *56*, 349-371.

Arnon, I., McCauley, S.M. & Christiansen, M.H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. *Journal of Memory and Language*, *92*, 265-280.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of- acquisition affects what gets learned. *Cognition, 122*, 292-305.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multiword phrases. *Journal of Memory and Language*, *62*, 67–82. Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*, *56*, 329-347.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438-481.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational Linguistics-Volume 1* (pp. 86-90). Association for Computational Linguistics.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*, 17284–17289.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, *19*, 241.

Bannard, C., & Ramscar, M. (2007). Reading time evidence for storage of frequent multiword sequences. Abstract in *Proceedings of the Architectures and Mechanism of Language Processing Conference (AMLAP-2007), Turku, Finland*.

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition*, *21*, 477–487.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing* (pp. 3–76). New York: Cambridge University Press.

Beckman, M. E., & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. E. Beckman (Eds.), *Between the grammar and physics of speech: Papers in laboratory phonology I* (pp. 152–178). Cambridge, UK: Cambridge University Press.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*, 92–111.

Berman, R. (1982). Verb-pattern alternation: The interface of morphology, syntax, and semantics in Hebrew child language. *Journal of Child Language, 9*, 169–191

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, *33*, 752–793.

Borensztajn, G., Zuidema, W., & Bod, R. (2009). Children's grammars grow more abstract with age: Evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science, 1*, 175–188.

Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In E. Wanner, & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 319-346). New York: Academic Press.

Brandt, S., Diessel, H., & Tomasello, M. (2008). The acquisition of German relative clauses: A case study. *Journal of Child Language, 35*, 325.

Brandt, S., Kidd, E., Lieven, E., & Tomasello, M. (2009). The discourse bases of relativization: An investigation of young German and English-speaking children's comprehension of relative clauses. *Cognitive Linguistics, 20*, 539-570.

Brandt, S., Lieven, E., & Tomasello, M. (2010). Development of word order in German complement-clause constructions: Effects of input frequencies, lexical items, and discourse function. *Language, 86*, 583-610.

Caldwell-Harris, C.L., Berant, J.B., & Edelman, S. (2012). Measuring mental entrenchment of phrases with perceptual identification, familiarity ratings, and corpus frequency statistics. In S. T. Gries & D. Divjak (Eds.), *Frequency effects in cognitive linguistics (Vol. 1): Statistical effects in learnability, processing and change.* The Hague, The Netherlands: De Gruyter Mouton.

Carlson, G. N., & Tanenhaus, M. K. (1988). Thematic roles and language comprehension. *Syntax and semantics, 21*, 263-288.

Carroll, S. (1989). Second-language acquisition and the computational paradigm. *Language Learning, 39*, 535-594.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*, 234–272.

Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research, 9*, 198–213.

Chang, N. C. L. (2008). *Constructing grammar: A computational model of the emergence of early constructions.* Unpublished doctoral dissertation. University of California, Berkeley.

Chater, N. & Christiansen, M.H. (2010). Language acquisition meets language evolution. *Cognitive Science, 34*, 1131-1157.

Chater, N. & Christiansen, M.H. (2016). Squeezing through the Now-or-Never bottleneck: Reconnecting language processing, acquisition, change and structure. *Behavioral & Brain Sciences, 39,* e62.

Chater, N. & Christiansen, M.H. (in press). Language acquisition as skill learning. *Current Opinion in Behavioural Sciences.*

Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language, 89,* 244-254.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19-22.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Christiansen, M.H. (in press). Implicit-statistical learning: A tale of two literatures. *Topics in Cognitive Science.*

Christiansen, M.H. & Arnon, I. (2017). More than words: the role of multiword sequences in language learning and use. *Topics in Cognitive Science, 9,* 542-551.

Christiansen, M.H. & Chater, N. (2016a). *Creating language: Integrating evolution, acquisition, and processing*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (2016b). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39,* e62.

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning, 59,* 126–161.

Clancy, P. M., Lee, H., & Zoh, M. H. (1986). Processing strategies in the acquisition of relative clauses: Universal principles and language-specific realizations. *Cognition*, *24*, 225–262.

Cohen, L., &Mehler, J. (1996). Click monitoring revisited: An on-line study of sentence

comprehension. *Memory & Cognition, 24*, 94–102.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Culicover, P. W. (2013). The role of linear order in the computation of referential dependencies. *Lingua, 136*, 125-144.

Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax*. New York: Oxford University Press.

Culicover, P. W., Jackendoff, R., & Audring, J. (2017). Multiword constructions in the grammar. *Topics in Cognitive Science, 9*, 552–568.

Dabrowska, E. (2000). From formula to schema: The acquisition of English questions. *Cognitive Linguistics, 11*, 83-102.

Demuth, K. 1992. *Acquisition of Sesotho*. In D. Slobin (ed.), *The cross-linguistic study of language acquisition, Vol. 3* (pp. 557-638). Hillsdale, N.J.: Lawrence Erlbaum Associates.

de Villiers, J. G., Tager-Flusberg, H. B., Hakuta, K., & Cohen, M. (1979). Children's comprehension of relative clauses. *Journal of Psycholinguistic Research, 8*, 499–518.

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge, UK: Cambridge University Press.

Diessel, H., & Tomasello, M. (2000). The development of relative clauses in spontaneous child speech. *Cognitive Linguistics, 11*, 131–152.

Diessel, H., & Tomasello, M. (2001). The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics*, *12*, 97–141.

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, *81*, 882–906.

Durrant, P. (2013). Formulaicity in an agglutinating language: The case of Turkish. *Corpus Linguistics and Linguistic Theory*, 9, 1-38.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 41, 375-396.

Ellis S. H. (1973). Structure and Experience in the Matching and Reproduction of Chess Patterns. Doctoral dissertation, Carnegie Mellon University, Pittsburgh.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

Edelman, S. (2008). *Computing the mind: how the mind really works*. Oxford University Press.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*, 540–551.

Feigenbaum, E. A., & Simon, H. A. (1962). A theory of the serial position effect. *British Journal of Psychology*, *53*, 307-320.

Ferreira, F. (2003). The misinterpretation of non-canonical sentences. *Cognitive Psychology*, *47*, 164–203.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*, 11.

Ferreira, F., & Patson, N. D. (2007). The "good enough" approach to language comprehension. *Language and Linguistics Compass, 1*, 71–83.

Fillenbaum, S. (1974). Pragmatic normalization: Further results for some conjunctive and disjunctive sentences. *Journal of Experimental Psychology*, *102*, 574-578.

Fisher, C., & Tokura, H. (1996). Acoustic cues to grammatical structure in infant-directed speech:

Cross-linguistic evidence. *Child Development*, *67*, 3192–3218.

Fitz, H., & Chang, F. (2008). The role of the input in a connectionist model of the accessibility hierarchy in development. In *Proceedings of the 32nd Boston University Conference on Language Development* (pp. 120-131).

Fitz, H. & Chang, F. (2017). Meaningful questions: The acquisition of auxiliary inversion in a connectionist model of sentence production. *Cognition*, *166*, 225-250.

Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*, 203–218.

Frank, S.L. & Christiansen, M.H. (in press). Hierarchical and sequential processing of language. *Language, Cognition and Neuroscience*.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*, 829.

Frank, S.L., Bod, R. & Christiansen, M.H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, *297*, 4522-4531.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107-125.

Frauenfelder, U., Segui, J., &Mehler, J. (1980). Monitoring around the relative clause. *Journal of Verbal Learning and Verbal Behavior*, *19*, 328–337.

Frazier, L. (1985). Syntactic complexity. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.) *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge, UK: Cambridge University Press.

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, *29*, 181–200.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review, 118*, 614.

Freudenthal, D., Pine, J. M., & Gobet, F. (2006). Modeling the development of children's use of optional infinitives in Dutch and English using MOSAIC. *Cognitive Science, 30*, 277-310.

Freudenthal, D., Pine, J. M., & Gobet, F. (2007). Understanding the developmental dynamics of subject omission: The role of processing limitations in learning. *Journal of Child Language, 34*, 83.

Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1366-1383.

Gagliardi, A., Mease, T., & Lidz, J. (submitted). *U-shaped development in the acquisition of filler-gap dependencies: Evidence from 15-and 20-month olds*.

Gathercole, V., Sebastian, E., & Soto, P. (1999). The early acquisition of Spanish verbal morphology: Across-the-board or piecemeal knowledge? *International Journal of Bilingualism, 3*, 138–182

Gertner, Y., & Fisher, C. (2012). Predicted errors in children's early sentence comprehension. *Cognition, 124,* 85-94.

Gervain, J., & Guevara Erra, R. (2012). The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition, 125*, 263-287.

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, *115*, 43-67.

Gobet, F., Freudenthal, D., & Pine, J. M. (2004). Modelling syntactic development in a cross-linguistic context. *Proceedings of the First Workshop on Psycho-computational Models of Human Language Acquisition* (pp. 53–60).

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*, 447-474.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. New York:

Oxford University Press.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American linguistics*, *26*, 178–194.

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2017). Facilitatory effects of multi-word units in lexical processing and word learning: A computational investigation. *Frontiers in Psychology, 8:555*. doi:10.3389/fpsyg.2017.00555

Hamburger, H., & Crain, S. (1982). Relative acquisition. *Language development*, *1*, 245–274.

Hammerton, J., Osborne, M., Armstrong, S., & Daelemans, W. (2002). Introduction to special issue on machine learning approaches to shallow parsing. *The Journal of Machine Learning Research*, *2*, 551–558.

Hamrick, P. (2014). A role for chunk formation in statistical learning of second language syntax. *Language Learning*, *64*, 247-278.

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2005). *The world atlas of linguistic structures*. Oxford, UK: Oxford University Press.

Hirsh-Pasek, K., Kemler Nelson, D. G., Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*, 269–286.

Holmes, V. M., & de la Bâtie, B. D. (1999). Assignment of grammatical gender by native speakers and foreign learners of French. *Applied Psycholinguistics, 20*, 479-506.

Holmes, V. M., & O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior*, *20*, 417–430.

Jackendoff, R. (1995). The boundaries of the lexicon. *Idioms: Structural and Psychological Perspectives*, 133–165.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* Oxford, UK: Oxford University Press.

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PLoS ONE, 7, e33202*.

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition, 13,* 789-845.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological bulletin, 131,* 684.

Jusczyk, P. W., Hirsh-Pasek, K., Kemler Nelson, D. G., Kennedy, L. J., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology, 24,* 252–293.

Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (in press). Meaning overrides frequency in idiomatic and compositional multiword chunks. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Jones, G. (2012). Why chunking should be considered as an explanation for developmental change before short-term memory capacity and processing speed. *Frontiers in Psychology, 3,* 167. doi: 10.3389/fpsyg.2012.00167.

Jones, G., Gobet, F., & Pine, J. M. (2000). A process model of children's early verb use. In L. R. Gleitman & A. K. Joshi (Eds.) *Proceedings of the 22nd Meeting of the Cognitive Science*

Society (pp. 723–728). Mahwah, NJ: Lawrence Erlbaum Associates.

Keenan, E. L., & Hawkins, S. (1987). The psychological validity of the accessibility hierarchy. *Universal Grammar*, *15*, 60–85.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*, 580–602.

Kipper-Schuler, K. (2006) *VerbNet: A broad-coverage, comprehensive verb lexicon.* Ph.D. thesis, University of Pennsylvania.

Kirjavainen, M., Theakston, A., & Lieven, E. (2009). Can input explain children's me-for-I errors? *Journal of Child Language, 36*, 1091-1114.

Klahr, D., Chase, W. G., & Lovelace, E. A. (1983). Structure and process in alphabetic retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*. 9, 462–477

Koh, S., Sanford, A. J., Clifton, C., & Dawydiak, E. J. (2008). Good-enough representation in plural and singular pronominal reference: Modulating the conjunction cost. In J. Gundel & N. Hedberg (Eds.) *Reference: Interdisciplinary Perspectives* (pp. 123-139). Oxford: Oxford University Press.

Kol, S., Nir, B., &Wintner, S. (2014). Computational evaluation of the Traceback Method. *Journal of Child Language, 41*, 176-199.

Kolodny, O., Lotem, A., & Edelman, S. (2015). Learning a generative probabilistic grammar of experience: A process-level model of language acquisition. *Cognitive Science*, *39*, 227-267.

Konopka, A. E., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, *58*, 68–101.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30*, 333–370.

MacDonald, M.C. & Christiansen, M.H. (2002). Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review, 109,* 35-54.

MacKay, D. G. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review, 89,* 483-506.

MacWhinney, B. (1975). Pragmatic patterns in child syntax. *Stanford Papers and Reports on Child Language Development, 10*, 153-165.

MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development, 43*, 1-123.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mandel, D. R., Jusczyk, P. W., & Kemler Nelson, D. G. (1994). Does sentential prosody help infants organize and remember speech information? *Cognition, 53,* 155–180.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT press.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development, 57 (4, Serial No. 228).*

Mariscal, S. (2008). Early acquisition of gender agreement in the Spanish noun phrase: starting small. *Journal of Child Language, 35*, 1-29.

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific Reports, 8,* 1079.

Maslen, R. J., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2004). A dense corpus study of past

tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research, 47,* 1319.

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1619-1624). Austin, TX: Cognitive Science Society.

McCauley, S. M., & Christiansen, M. H. (2013). Toward a unified account of comprehension and production in language development. *Behavioral and Brain Sciences, 36,* 366-367.

McCauley, S.M. & Christiansen, M.H. (2014a). Acquiring formulaic language: A computational model. *Mental Lexicon, 9,* 419-436.

McCauley, S.M. & Christiansen, M.H. (2014b). Prospects for usage-based computational models of grammatical development: Argument structure and semantic roles. *Wiley Interdisciplinary Reviews: Cognitive Science, 5,* 489-499.

McCauley, S.M. & Christiansen, M.H. (in press). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science.*

McCauley, S.M. & Christiansen, M.H. (2015). Individual differences in chunking ability predict on-line sentence processing. In D.C. Noelle & R. Dale (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

McCauley, S.M., Monaghan, P. & Christiansen, M.H. (2015). Language emergence in development: A computational perspective. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 415-436). Hoboken, NJ: Wiley-Blackwell.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review, 63*, 81.

Miller, G. A. (1958). *Free recall of redundant strings of letters. Journal of Experimental Psychology,*

56, 485–491.

Miller, G. A. (1990). Nouns in WordNet: a lexical inheritance system. *International journal of Lexicography, 3*, 245-264.

Miller, G. A., & Taylor, W. G. (1948). The perception of repeated bursts of noise. *The Journal of the Acoustical Society of America*, *20*, 171-182.

Miyata, S. (2000). The TAI corpus: Longitudinal speech data of a Japanese boy aged 1;5.20–3;1.1. *Bulletin of Shukutoku Junior College, 39*, 77–85.

Miyata, S., & Naka, N. (1998). Wakachigaki Guideline for Japanese: WAKACHI98 v.1.1. The Japanese Society for Educational Psychology Forum Report No. FR-98-003, The Japanese Association of Educational Psychology.

Monaghan, P. & Christiansen, M.H. (2008). Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data (TILAR Series)* (pp. 139-163). Amsterdam: John Benjamins.

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*, 545-564.

O'Grady, W. (2015). Anaphora and the case for emergentism. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 100-122). Hoboken, NJ: Wiley-Blackwell.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*, 244–247.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, *36*, 1299-1305.

Perruchet, P., Poulin-Charronnat, B., Tillmann, B., & Peereman, R. (2014). New evidence for chunk-based models in word segmentation. *Acta Psychologica*, *149*, 1-8.

Perruchet, P., &Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*, 246-263.

Perruchet, P., Vinter, A., Pacteau, C., & Gallego, J. (2002). The formation of structurally relevant units in artificial grammar learning. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(2), 485-503.

Peters, A. M. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge University Press.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*, 105–110.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329-347.

Pickering, M. J. & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 10.1037/bul0000158.*

Pine, J. M., & Lieven, E. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics, 18*, 123-138.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Harper Collins.

Pizutto, E. & Caselli, C. (1992). The acquisition of Italian morphology. *Journal of Child Language, 19*, 491–557.

Punyakanok, V., & Roth, D. (2001). The use of classifiers in sequential inference. In *Proceedings of NIPS 2001* (pp. 995-1001).

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language, 89*, 760-793.

Reali, F. & Christiansen, M.H. (2007). Word-chunk frequencies affect the processing of pronominal

object-relative clauses. *Quarterly Journal of Experimental Psychology, 60,* 161-170.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6,* 855-863.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-469.

Remez, R. E., Ferro, D. F., Dubowski, K. R., Meer, J., Broder, R. S., & Davids, M. L. (2010). Is desynchrony tolerance adaptable in the perceptual organization of speech? *Attention, Perception, & Psychophysics*, *72*, 2054-2058.

van Rijsbergen, C. J. (1979). *Information Retrieval*. London: Butterworths.

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker: A MDL-based cognitive model of inductive learning. *Cognitive science, 35,* 1352-1389.

Rogers, M. (1987). Learners difficulties with grammatical gender in German as a foreign language. *Applied Linguistics, 8*, 48-74.

Rowland, C. F. (2007). Explaining errors in children's questions. *Cognition, 104,* 106-134.

Rubino, R. and Pine, J. (1998) Subject–verb agreement in Brazilian Portuguese: What low error rates hide. *Journal of Child Language, 25*, 35–60.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*, 493–515.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language, 47*, 172–196.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-621.

Sanford, A. J. (2002). Context, attention, and depth of processing during interpretation. *Mind and Language*, *17*, 188–206.

Sanford, A. J., & Garrod, S. C. (1981). *Understanding written language: Explorations of comprehension beyond the sentence*. New York: Wiley.

Sanford, A. J., & Garrod, S. C. (1998). The role of scenario mapping in text comprehension. *Discourse Processes*, *26*, 159–190.

Sanford, A. J. S., Sanford, A. J., Filik, R., & Molle, J. (2005). Depth of lexical-semantic processing and sentential load. *Journal of Memory and Language*, *53*, 378–396.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*, 382–386.

Schiffman, H. F. (1999). *A reference grammar of spoken Tamil.* Cambridge, UK: Cambridge University Press.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. Proceedings of the ACL SIGDAT-Workshop, March 1995.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592.

Simon, H. A. (1974). How big is a chunk? *Science, 183,* 482-488.

Simon H. A. & Gilmartin K. J. (1973). A simulation of memory for chess positions. Cognitive Psychology, 5, 29-46.

Simon D. P. & Simon H. A. (1973). Alternative uses of phonemic information in spelling. Rev. Educ. Res. 43, 115–137.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, *5*, 644–649.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-

tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27,* 251-272.

Slobin, D. I. (1986). *The crosslinguistic study of language acquisition: The data (Vol. 2).* London: Psychology Press.

Soderstrom, M., Seidl, A., Kemler Nelson, D. G., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, *49*, 249–267.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, *102*, 11629-11634.

Sprenger, S. A., Levelt, W. J., &Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, *54*, 161–184.

Stemberger, J.P., & Bernhardt, B.H., & Johnson, C.E. (1999). "Regressions" ("u"-shaped learning) in the acquisition of prosodic structure. Poster presented at the 6th International Child Language Congress, July 1999.

Studdert-Kennedy, M. (1986). Some developments in research on language behavior. *Behavioral and Social Science: 50 Years of Discovery*, 208.

Sturt, P., Sanford, A. J., Stewart, A., &Dawydiak, E. (2004). Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin & Review*, *11*, 882–888.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*, *36*, 201-216.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*, 86–132.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language, 50*, 355-370.

Tanenhaus, M. K., Carlson, G., & Trueswell, J. C. (1989). The role of thematic structures in interpretation and parsing. *Language and Cognitive Processes, 4*, 211-234.

Tavakolian, S. L. (1977). *Structural principles in the acquisition of complex sentences*. Unpublished doctoral dissertation. University of Massachusetts, Amherst.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*, 1-42.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, US: Harvard University Press.

Tomasello, M. and Brooks, P. (1998) Young children's earliest transitive and intransitive constructions. *Cognitive Linguistics, 9,* 379–395.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.) *Perspectives on formulaic language: Acquisition and communication* (pp. 151-173). London: Continuum International Publishing Group.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing Advantages of Lexical Bundles: Evidence from self‐paced reading and sentence recall tasks. *Language Learning, 61*, 569-613.

Tsarfaty, R., Nivre, J., &Anderson, E. (2012). Joint evaluation of morphological segmentation and syntactic parsing. In *Proceedings of the 50th Annual Meeting of the Association for*

Computational Linguistics: Short Papers - Volume 2 (pp. 6-10). Association for Computational Linguistics.

Tunstall, S. L. (1998). *The interpretation of quantifiers: Semantics and processing*. Unpublished doctoral dissertation. University of Massachusetts, Amherst.

Tyler, L. K., & Marslen-Wilson, W. D. (1977). The on-line effects of semantic context on syntactic processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 683-692.

Wanner, E. & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan, & G. A. Miller (Eds.) *Linguistic theory and psychological reality* (pp. 119-161). Boston: MIT Press.

Warren, R. M., Obusek, C. J., Farmer, R. M., & Warren, R. P. (1969). Auditory sequence: Confusion of patterns other than speech or music. *Science, 164*, 586-587.

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *The Quarterly Journal of Experimental Psychology*, *31*, 591–597.

Wray, A. (2005). *Formulaic language and the lexicon.* Cambridge, UK: Cambridge University Press.

Appendix A: Shallow parsing accuracy and completeness statistics for English, French, and German

English

In line with the developmental motivation for the model, we examined accuracy rates independently. Across the 43 child corpora, CBL achieved a mean accuracy rate of 76.4%, while PARSER attained a mean accuracy of 65.2% and the Trigram model reached a mean accuracy rate of 65.8%. The scores are shown in Table A1. As can be seen, CBL not only outperformed its baselines, but once more yielded a tighter, more uniform distribution of scores

As in our analysis of the overall comprehension performance scores, we submitted the logit-transformed accuracy scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram) with *Child Corpus* as a random factor. This yielded a significant main effect of *Model* [$F(2,84) = 1877$, $p < 0.0001$], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [$t(42)=65.17$, $p<0.0001$] and Trigram [$t(42)=39$, $p<0.0001$] models, as well as stronger performance for the Trigram model compared to PARSER [$t(42)=2.63$, $p<0.05$].

Finally, we looked at completeness scores. Across the 43 child corpora, CBL achieved a mean completeness of 73.8%, while the PARSER attained a mean completeness of 68.7% and the Trigram model reached a mean completeness rate of 66.5%. The scores are shown in Table A1.

As with accuracy, we submitted the logit-transformed completeness scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram) with *Child Corpus* as a random factor. This yielded a significant main effect of *Model* [$F(2,84) = 42.14$, $p < 0.0001$], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [$t(42)=7.77$,

p<0.001] and Trigram [t(42)=11.9, p<0.0001] models, with no significant difference in means for PARSER relative to the Trigram model [t(42)=1.94, p=0.06].

Table A1
Shallow Parsing Accuracy and Completeness for English

	Accuracy	**Completeness**
CBL	76.4%	73.8%
PARSER	65.2%	68.7%
Trigram	65.8%	66.5%

French

As with the English simulations, we examined accuracy separately. Across the 15 child corpora, CBL attained a mean accuracy rate of 72.0%, while the PARSER model attained a mean accuracy rate of 61.8%. The Trigram model attained a mean accuracy rate of 57.0%. The scores are shown in Table A1.

As with the previous analyses, we submitted the logit-transformed accuracy scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(2,26) = 342.3, p < 0.0001], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [t(14)=24.54, p<0.0001] and Trigram [t(14)=18.69, p<0.0001] models, as well as for PARSER compared to the Trigram model [t(14)=9.7, p=0.0001].

We also analyzed completeness: across the 15 child corpora, CBL attained a mean completeness score of 70.8%, while the PARSER model attained a mean completeness rate of 73.5%. The Trigram model attained a mean completeness rate of 66.1%. The scores are shown in Table A2. As with accuracy, we submitted the logit-transformed completeness scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(2,26) = 21.96, p < 0.0001], with post-hoc analyses confirming stronger performance for PARSER compared to the CBL [t(14)=2.54, p<0.05] and

Trigram [t(14)=5.29, p<0.001] models, as well as for CBL compared to the Trigram model [t(14)=6.35, p=0.0001].

Table A2
Shallow Parsing Accuracy and Completeness for German

	Accuracy	**Completeness**
CBL	72.0%	70.8%
PARSER	61.8%	73.5%
Trigram	57.0%	66.1%

German

As with the English and French simulations, we examined German accuracy separately. Across the 22 child corpora, CBL attained a mean accuracy rate of 78.0%, while PARSER attained a mean accuracy rate of 69.4%. The Trigram model attained an accuracy of 70.5%. The scores are shown in Table A3.

We once more submitted the logit-transformed accuracy scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(2,40) = 475.2, p < 0.0001], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [t(21)=29.4, p<0.0001] and Trigram [t(21)=20.05, p<0.0001] models, as well as for the Trigram model compared to PARSER [t(21)=4.23, p=0.001].

As with the English and French simulations, we also examined completeness separately. Across the 22 child corpora, CBL attained a mean completeness of 72.2%, while PARSER attained a mean completeness of 83.5%. The Trigram model attained a completeness of 62.9%. The scores are shown in Table A3.

We once more submitted the logit-transformed completeness scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(2,40) = 61.7, p < 0.0001], with post-hoc

analyses confirming stronger performance for PARSER compared to the CBL [$t(21)=5.49$, $p<0.0001$] and Trigram [$t(21)=8.59$, $p<0.0001$] models, as well as for CBL compared to the Trigram model [$t(14)=11.3$, $p=0.0001$].

Table A3
Shallow Parsing Accuracy and Completeness for German

	Accuracy	**Completeness**
CBL	78.0%	72.2%
PARSER	69.4%	83.5%
Trigram	70.5%	62.9%

Appendix B: Examples of Frequent Chunks

Table B1: Most Frequent Chunks Across Three Timesteps: English Dense Corpus Simulation

1000 Utterances	10,000 Utterances	100,000 Utterances
oh dear	*oh dear*	*oh dear*
is it	*what's this*	*I think*
the giraffe	*I think*	*all done*
a hat	*you like*	*what's this*
I think	*look at*	*that's right*
the basket	*that's right*	*and then*
the steps	*is it*	*isn't it*
the lion	*all done*	*is it*
what does	*and there's*	*is that*
what do	*oh dear dear*	*look at*
are they	*this morning*	*you like*
it's gone	*and then*	*you see*
you like	*what does*	*a big*
and there's	*going to*	*you want*
the wind	*the bus*	*you can*
what about under	*a ride*	*what do*
the boat	*what do*	*you've got*
on mummy's head	*the door*	*on the floor*
that's Thomas	*a hat*	*a mess*
your birthday cards	*as well*	*bye bye*

Table B2: Most Frequently Chunked Items by Category: English Dense Corpus Simulation

1000 Utterances	**10,000 Utterances**	**100,000 Utterances**
DET NOUN	*DET NOUN*	*DET NOUN*
PRO VERB	*PRO VERB*	*PRO VERB*
NOUN VERB	*NOUN VERB*	*NOUN VERB*
ADJ NOUN	*PRO NOUN*	*ADJ NOUN*
DET ADJ NOUN	*PREP NOUN*	*PREP NOUN*
NOUN ADV	*ADJ NOUN*	*DET ADJ NOUN*
CONJ NOUN	*DET ADJ NOUN*	*PRO NOUN*
INTRJ NOUN	*VERB NOUN*	*PREP DET NOUN*
NOUN PREP	*CONJ NOUN*	*VERB NOUN*
VERB PRO	*NUM NOUN*	*CONJ NOUN*

Note: ADJ = adjective; ADV = adverb; CONJ = conjunction; DET = determiner; INTRJ = interjection; NOUN = noun; NUM = numeral; PREP = preposition; PRO = pronoun; VERB = verb.

Appendix C: Evaluating the Effects of Forwards vs. Backwards Transitional Probability

Baseline Models

We created three baseline models in order to explore a 2 x 2 design, depicted in Table C1, including the factors *unit type* (chunks vs. *n*-grams) and *direction* (backward vs. forward transitional probability).

Table C1
Contrasting Direction and Unit Type

	Chunks	**N-grams**
BTP	CBL	BTP3G
FTP	FTP-Chunk	FTP3G

As previous work in the statistical learning literature has focused on FTP as a cue to phrase structure (e.g., Thompson & Newport, 2007), an alternate model was created to compare the usefulness of this cue against the BTPs used by CBL. Thus, the first baseline model, hereafter referred to as the FTP-Chunk model, was identical to CBL, with the sole exception that all BTP calculations were replaced by FTP calculations.

As the Trigram model described in the main paper relied on FTPs, we created an otherwise identical baseline model which relied on BTP rather than FTP calculations. Both models learned trigram statistics in an incremental, on-line fashion, in the style of CBL, while simultaneously processing utterances through the placement of chunk boundaries. In the present appendix we refer to the Trigram baseline as the FTP3G baseline, and the backwards transitional probability version as the BTP3G baseline.

In the case of the FTP3G baseline, if the FTP between the first bigram and the final unigram of a trigram fell below the running average for the same statistic, a chunk boundary was inserted. For

instance, as the model encountered Z after seeing the bigram XY, it would calculate the FTP for the trigram by normalizing the frequency count of the trigram XYZ by the count of the bigram XY, and comparing the result to the running average FTP for previously encountered trigrams (inserting a chunk boundary if the running average was greater). In the case of the BTP3G baseline, a chunk boundary was placed if the BTP between the first unigram and the final bigram of the trigram fell below the running average. The start-of-utterance marker made it possible for the 3G baselines to place a boundary between the first and second words of an utterance. During production attempts, which were also incremental and on-line in the style of CBL, both trigram models began constructing an utterance by choosing from the bag-of-words the word with the highest TP (FTP for the FTP3G model, and BTP for the BTP3G model), given the start-of-utterance marker (in other words, bigram statistics were used to select the first word). Each subsequent word was chosen according to trigram statistics, based on the two most recently placed words (or the initial word and the start-of-utterance marker, in the case of selecting the second word in an utterance). For the FTP3G model, this meant the word with the highest FTP given the two preceding words was chosen; for the BTP3G model, the word resulting in the highest BTP between the final bigram and the first unigram of the resulting trigram was chosen. Thus, like CBL and its FTP-based counterpart, both trigram baseline models relied on identical statistics during comprehension and production (either BTPs or FTPs, computed over trigrams).

Shallow Parsing Results

Shallow parsing results for the same English child corpora are shown for the model and its baselines in Table C2.

Table C2: English Shallow Parsing F-Scores

	Chunks	**N-grams**

BTP	75.4	61.3
FTP	67.5	65.9

We submitted the shallow parsing F-scores (logit-transformed) to a repeated-measures ANOVA with the factors *Unit Type* (2: Chunks vs. *n*-grams) and *Direction* (2: BTP vs. FTP), with *Child Corpus* as a random factor. This yielded main effects of *Unit Type* [$F(1,42) = 1184$, $p < 0.0001$] and *Direction* [$F(1,42) = 78.01$, $p < 0.0001$], indicating better performance for chunk-based models and BTPs, respectively, and a significant *Unit Type* x *Direction* interaction [$F(1,42) = 792.5$, $p < 0.0001$], indicating better performance for the CBL model's combination of BTPs and chunks.

The French shallow parsing scores, depicted in Table C3, followed the same qualitative pattern as the English data.

Table C3: French Shallow Parsing F-Scores

	Chunks	**N-grams**
BTP	71.6	51.6
FTP	61.7	59.0

We submitted the French shallow parsing F-scores (logit-transformed) to a repeated-measures ANOVA with the factors *Unit Type* (2: Chunks vs. *n*-grams) and *Direction* (2: BTP vs. FTP), with *Child Corpus* as a random factor. This yielded main effects of *Unit Type* [$F(1,14) = 573.5$, $p < 0.0001$] and *Direction* [$F(1,14) = 14.7$, $p < 0.01$], indicating better performance for chunk-based models and BTPs, respectively, and a significant *Unit Type* x *Direction* interaction [$F(1,14) = 234.2$, $p < 0.0001$], indicating better performance for the CBL model's combination of BTPs and chunks.

The German shallow parsing scores, shown in Table C4, followed once more the same general pattern.

Table C4: German Shallow Parsing F-Scores

	Chunks	**N-grams**
BTP	75.7	53.2
FTP	71.0	67.4

We submitted the German shallow parsing F-scores (logit-transformed) to a repeated-measures ANOVA with the factors *Unit Type* (2: Chunks vs. *n*-grams) and *Direction* (2: BTP vs. FTP), with *Child Corpus* as a random factor. This yielded main effects of *Unit Type* [$F(1,203) = 890.2$, $p < 0.0001$] and *Direction* [$F(1,203) = 47.5$, $p < 0.0001$], indicating better performance for chunk-based models and BTPs, respectively, and a significant *Unit Type* x *Direction* interaction [$F(1,203) = 389.8$, $p < 0.0001$], indicating better performance for the CBL model's combination of BTPs and chunks.

Sentence Production Performance Results

The sentence production performance results for the CBL model, the FTP-Chunk model, and the two 3G baselines are shown in Figure C1.

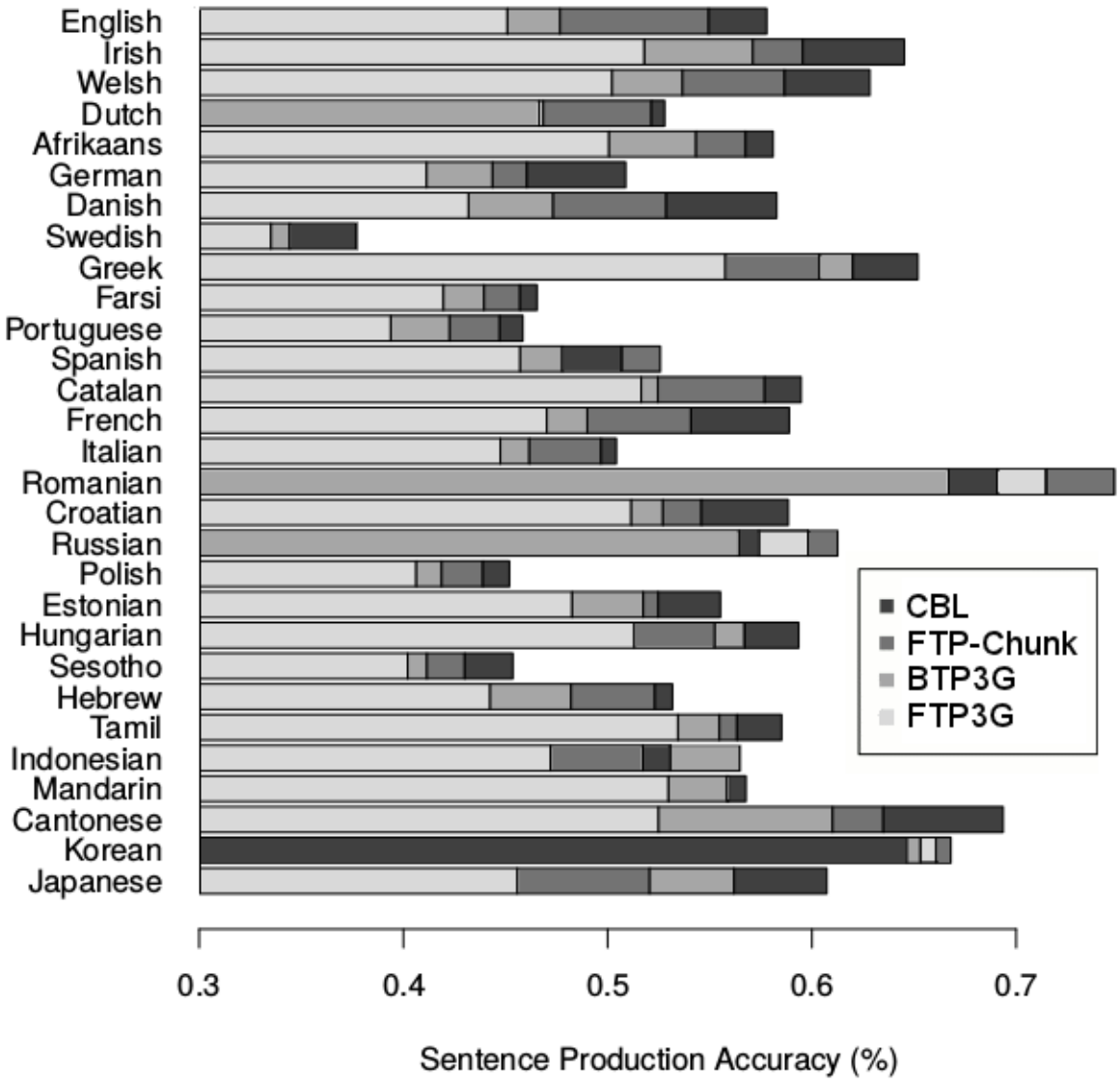


Fig. C1: Mean Sentence Production Accuracy scores for the CBL model and its trigram baseline across all 29 languages, including English (shown at top). Bars are non-cumulative

We submitted the sentence production performance F-scores (logit-transformed) to a repeated-measures ANOVA with the factors *Unit Type* (2: Chunks vs. *n*-grams) and *Direction* (2: BTP vs. FTP), with *Child Corpus* as a random factor. This yielded main effects of *Unit Type* [$F(1,21) = 801.4$, $p < 0.0001$] and *Direction* [$F(1,21) = 128.4$, $p < 0.0001$], indicating better performance for chunk-based models and BTPs, respectively.

Appendix D: Corpus List and Citations

Language	Child	Citation
Afrikaans	Chan	No citation provided
Afrikaans	Jean	No citation provided
Cantonese	Chunyat	Lee, T. H.T., Wong, C. H., Leung, S., Man. P., Cheung, A., Szeto, K., & Wong, C. S. P. (1996). The development of grammatical competence in Cantonese-speaking children (RGC Project No. CUHK 2/91). Hong Kong: Hong Kong Research Grant Committee.
Cantonese	Gakei	Lee, T. H.T., Wong, C. H., Leung, S., Man. P., Cheung, A., Szeto, K., & Wong, C. S. P. (1996). The development of grammatical competence in Cantonese-speaking children (RGC Project No. CUHK 2/91). Hong Kong: Hong Kong Research Grant Committee.
Cantonese	Tsuntsun	Lee, T. H.T., Wong, C. H., Leung, S., Man. P., Cheung, A., Szeto, K., & Wong, C. S. P. (1996). The development of grammatical competence in Cantonese-speaking children (RGC Project No. CUHK 2/91). Hong Kong: Hong Kong Research Grant Committee.
Cantonese	Johnny	Lee, T. H.T., Wong, C. H., Leung, S., Man. P., Cheung, A., Szeto, K., & Wong, C. S. P. (1996). The development of grammatical competence in Cantonese-speaking children (RGC Project No. CUHK 2/91). Hong Kong: Hong Kong Research Grant Committee.
Cantonese	Jenny	Lee, T. H.T., Wong, C. H., Leung, S., Man. P., Cheung, A., Szeto, K., & Wong, C. S. P. (1996). The development of grammatical competence in Cantonese-speaking children (RGC Project No. CUHK 2/91). Hong Kong: Hong Kong Research Grant Committee.
Cantonese	Tinfaan	Lee, T. H.T., Wong, C. H., Leung, S., Man. P., Cheung, A., Szeto, K., & Wong, C. S. P. (1996). The development of grammatical competence in Cantonese-speaking children (RGC Project No. CUHK 2/91). Hong Kong: Hong Kong Research Grant Committee.
Cantonese	Bernard	Lee, T. H.T., Wong, C. H., Leung, S., Man. P., Cheung, A., Szeto, K., & Wong, C. S. P. (1996). The development of grammatical competence in Cantonese-speaking children (RGC Project No. CUHK 2/91). Hong Kong: Hong Kong Research Grant Committee.
Cantonese	Bohuen	Lee, T. H.T., Wong, C. H., Leung, S., Man. P., Cheung, A., Szeto, K., & Wong, C. S. P. (1996). The development of grammatical competence in Cantonese-speaking children (RGC Project No. CUHK 2/91). Hong Kong: Hong Kong Research Grant Committee.
Catalan	Gisela	No citation provided
Catalan	Guillem	No citation provided
Catalan	Jordina	No citation provided
Catalan	Laura	No citation provided
Croatian	Antonija	Kovacevic, M. (2003). Acquisition of Croatian in crosslinguistic perspective. Zagreb.

Croatian	Marina	Kovacevic, M. (2003). Acquisition of Croatian in crosslinguistic perspective. Zagreb.
Croatian	Vjeran	Kovacevic, M. (2003). Acquisition of Croatian in crosslinguistic perspective. Zagreb.
Danish	Anne	Plunkett, K. (1985). Preliminary approaches to language development. Århus: Århus University Press.
Danish	Jens	Plunkett, K. (1985). Preliminary approaches to language development. Århus: Århus University Press.
Dutch	Abe	Wijnen, F. & M. Verrips (1998). The acquisition of Dutch syntax. In S. Gillis & A. De Houwer (Eds.), *The acquisition of Dutch.* Amsterdam: John Benjamins.
Dutch	Arnold	Schaerlaekens, A. M. (1973). The two-word sentence in child language. The Hague: Mouton.
Dutch	Daan	Wijnen, F. & M. Verrips (1998). The acquisition of Dutch syntax. In S. Gillis & A. De Houwer (Eds.), *The acquisition of Dutch.* Amsterdam: John Benjamins.
Dutch	Iris	Wijnen, F. & M. Verrips (1998). The acquisition of Dutch syntax. In S. Gillis & A. De Houwer (Eds.), *The acquisition of Dutch.* Amsterdam: John Benjamins.
Dutch	Josse	Wijnen, F. & M. Verrips (1998). The acquisition of Dutch syntax. In S. Gillis & A. De Houwer (Eds.), *The acquisition of Dutch.* Amsterdam: John Benjamins.
Dutch	Laura	van Kampen, J. (2009). The non-biological evolution of grammar: Wh-question formation in Germanic. *Biolinguistics, 2*, 154-185.
Dutch	Maarten	Schaerlaekens, A., & Gillis, S. (1987). De taalverwerving van het kind: Eenhernieuwdeorientatie in het Nederlandstaligonderzoeks. Groningen: Wolters-Noordhoff.
Dutch	Matthijs	Wijnen, F. & M. Verrips (1998). The acquisition of Dutch syntax. In S. Gillis & A. De Houwer (Eds.), *The acquisition of Dutch.* Amsterdam: John Benjamins.
Dutch	Niek	Elbers, L., & Wijnen, F. (1993). Effort, production skill, and language learning. In C. Stoel-Gammon (Ed.) *Phonological development.* Timonium, MD: York.
Dutch	Peter	Wijnen, F. & M. Verrips (1998). The acquisition of Dutch syntax. In S. Gillis & A. De Houwer (Eds.), *The acquisition of Dutch.* Amsterdam: John Benjamins.
Dutch	Sarah	van Kampen, J. (2009). The non-biological evolution of grammar: Wh-question formation in Germanic. *Biolinguistics, 2*, 154-185.
Dutch	Tom	Wijnen, F. & M. Verrips (1998). The acquisition of Dutch syntax. In S. Gillis & A. De Houwer (Eds.), *The acquisition of Dutch.* Amsterdam: John Benjamins.
English	Abe	Kuczaj, S. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior, 16*, 589–600.
English	Adam	Brown, R. (1973). A first language: The early stages. Cambridge, MA: Harvard University Press.
English	Alex	Demuth, K. & McCullough, E. (2009). The prosodic (re)organization of children's early English articles. *Journal of Child Language, 36,* 173-200.
English	Anne	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	Aran	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.

English	Barbara	Henry, A. (1995). Belfast English and Standard English: Dialect variation and parameter setting. New York: Oxford University Press.
English	Becky	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	Carl	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	Conor	Henry, A. (1995). Belfast English and Standard English: Dialect variation and parameter setting. New York: Oxford University Press.
English	David	Henry, A. (1995). Belfast English and Standard English: Dialect variation and parameter setting. New York: Oxford University Press.
English	Dominic	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	Emily	Weist, R. M. & Zevenbergen, A. (2008). Autobiographical memory and past time reference. *Language Learning and Development, 4,* 291 – 308.
English	Emma	Weist, R. M. & Zevenbergen, A. (2008). Autobiographical memory and past time reference. *Language Learning and Development, 4,* 291 – 308.
English	Ethan	Demuth, K. & McCullough, E. (2009). The prosodic (re)organization of children's early English articles. *Journal of Child Language, 36,* 173-200.
English	Eve	Brown, R. (1973). A first language: The early stages. Cambridge, MA: Harvard University Press.
English	Gail	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	Jilly	Weist, R. M. & Zevenbergen, A. (2008). Autobiographical memory and past time reference. *Language Learning and Development, 4,* 291 – 308.
English	Jimmy	Demetras, M. (1989). Changes in parents' conversational responses: A function of grammatical development. Paper presented at ASHA, St. Louis, MO.
English	Joel	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	John	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	Lara	Rowland, C. F. & Fletcher, S. L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language, 33*, 859-877.
English	Lily	Demuth, K. & McCullough, E. (2009). The prosodic (re)organization of children's early English articles. *Journal of Child Language, 36,* 173-200.
English	Liz	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	Matt	Weist, R. M. & Zevenbergen, A. (2008). Autobiographical memory and past time reference. *Language Learning and Development, 4,* 291 – 308.

English	Michelle	Henry, A. (1995). Belfast English and Standard English: Dialect variation and parameter setting. New York: Oxford University Press.
English	Nai	Demuth, K. & McCullough, E. (2009). The prosodic (re)organization of children's early English articles. *Journal of Child Language, 36,* 173-200.
English	Naomi	Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent–child discourse. In K. E. Nelson (Ed.), *Children's language, Vol. 4* (pp. 1-28), Hillsdale, NJ: Lawrence Erlbaum Associates.
English	Nathaniel	MacWhinney, B., & Snow, C. (1990). The Child Language Data Exchange System: An update. *Journal of Child Language, 17,* 457-472.
English	Nic	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	Nina	Suppes, P. (1974). The semantics of children's language. *American Psychologist, 29,* 103– 114.
English	Peter	Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when and why. *Cognitive Psychology, 6*, 380–420.
English	Roman	Weist, R. M. & Zevenbergen, A. (2008). Autobiographical memory and past time reference. *Language Learning and Development, 4,* 291 – 308.
English	Ross	MacWhinney, B. (1991). The CHILDES project: Tools for analyzing talk. Hillsdale, NJ: Erlbaum.
English	Ruth	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	Sarah	Brown, R. (1973). A first language: The early stages. Cambridge, MA: Harvard University Press.
English	Seth	Peters, A. (1987). The role of imitation in the developing syntax of a blind child. *Text, 7,* 289–311.
English	Shem	Clark, E. V. (1978). Awareness of language: Some evidence from what children say and do. In R. J. A. Sinclair & W. Levelt (Eds.), *The child's conception of language* (pp. 17-43). Berlin: Springer Verlag.
English	Thomas	Lieven, E., Salomo, D. & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20*, 481-508.
English	Tow	Demetras, M., Post, K., & Snow, C. (1986). Feedback to first-language learners. *Journal of Child Language, 13*, 275–292.
English	Trevor	Demetras, M. (1989). Working parents conversational responses to their two-year-old sons. Working paper. University of Arizona.
English	Violet	Demuth, K. & McCullough, E. (2009). The prosodic (re)organization of children's early English articles. *Journal of Child Language, 36,* 173-200.
English	Warren	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.
English	Will	Demuth, K. & McCullough, E. (2009). The prosodic (re)organization of children's early English articles. *Journal of Child Language, 36,* 173-200.
Estonian	Henri	Kohler, K. (2004) Erwerb der frühen Verbmorphologieim Estnischen. Unpublished doctoral thesis, University of Potsdam.

Estonian	Hendrik	Argus, R. (1998). CHILDES'ieestiandmepank ja sellesuhtluskeskneanalüüs (Hendrik, 1.6-2.6). Magistritöö, TallinnaPedagoogikaülikool, filoloogiateaduskond, eestikeeleõppetool. Tallinn: Tallinna Pedagoogikaülikool.
Estonian	Antsu	No citation provided
Farsi	Lilia	Family, N. (2009). Lighten up: The acquisition of light verb constructions in Persian. In J. Chandlee, M. Franchini, S. Lord, & G-M. Rheiner (Eds.) *Proceedings of BUCLD 33*. Sommerville, MA: Cascadilla Press.
Farsi	Minu	Family, N. (2009). Lighten up: The acquisition of light verb constructions in Persian. In J. Chandlee, M. Franchini, S. Lord, & G-M. Rheiner (Eds.) *Proceedings of BUCLD 33*. Sommerville, MA: Cascadilla Press.
French	Anais	Demuth, K. & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners*. Journal of Child Language, 35*, 99-127.
French	Greg	Champaud, C. (1994). The development of verb forms in French children at around two years of age: some comparisons with Romance and non-Romance languages. Paper presented at the First Lisbon Meeting on Child Language, Lisbon, Portugal.
French	Leonard	Leroy, M., Mathiot, E., & Morgenstern, A. (2009). Pointing gestures and demonstrative words: Deixis between the ages of one and three. In J. Zlatev, M. J. Falck, C. Lundmark, & M. Andrén (Eds.) *Studies in language and cognition* (pp. 386-404). Cambridge: Cambridge Scholars Publishing.
French	Liea	De Cat, C. & Plunkett, B. (2002). 'Qu'est ce qu'i(l) dit, celui +la`?' Notes methodologiques sur la transcription d'un corpus francophone. In C. D. Pusch & W. Raible (eds), Romance corpus linguistics : Corpora and spoken language. Tubingen: Narr.
French	Madeleine	Leroy, M., Mathiot, E., & Morgenstern, A. (2009). Pointing gestures and demonstrative words: Deixis between the ages of one and three. In J. Zlatev, M. J. Falck, C. Lundmark, & M. Andrén (Eds.) *Studies in language and cognition* (pp. 386-404). Cambridge: Cambridge Scholars Publishing.
French	Marie	Hamann, C., Ohayon, S., Dubé, S., Frauenfelder, U. H., Rizzi, L., Starke, M., et al. (2003). Aspects of grammatical development in young French children with SLI. *Developmental Science, 6,* 151-160.
French	Marie	Demuth, K. & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners*. Journal of Child Language, 35*, 99-127.
French	Mona	De Cat, C. & Plunkett, B. (2002). 'Qu'est ce qu'i(l) dit, celui +la`?' Notes methodologiques sur la transcription d'un corpus francophone. In C. D. Pusch & W. Raible (eds), Romance corpus linguistics : Corpora and spoken language. Tubingen: Narr.
French	Nathan	Demuth, K. & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners*. Journal of Child Language, 35*, 99-127.
French	Para	De Cat, C. & Plunkett, B. (2002). 'Qu'est ce qu'i(l) dit, celui +la`?' Notes methodologiques sur la transcription d'un corpus francophone. In C. D. Pusch

		& W. Raible (eds), Romance corpus linguistics : Corpora and spoken language. Tubingen: Narr.
French	Pauline	Bassano, D. & Maillochon, I. (1994). Early grammatical and prosodic marking of utterance modality in French : a longitudinal case study. *Journal of Child Language, 21*, 649-675.
French	Phil	Suppes, P., Smith, R., & Leveillé, M. (1973). The French syntax of a child's noun phrases. *Archives de Psychologie, 42*, 207–269.
French	Rondal	Rondal, J. A. (1985). Adult–child interaction and the process of language understanding. New York: Praeger.
French	Theophile	Leroy, M., Mathiot, E., & Morgenstern, A. (2009). Pointing gestures and demonstrative words: Deixis between the ages of one and three. In J. Zlatev, M. J. Falck, C. Lundmark, & M. Andrén (Eds.) *Studies in language and cognition* (pp. 386-404). Cambridge: Cambridge Scholars Publishing.
French	Theotime	Demuth, K. & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners*. Journal of Child Language, 35*, 99-127.
German	Andreas	Wagner, K. R. (1985). How much do children say in a day? *Journal of Child Language, 12*, 475–487.
German	Ann	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21*, 109-141.
German	Caroline	No citation provided
German	Cosima	No citation provided
German	Emely	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21*, 109-141.
German	Falko	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21*, 109-141.
German	Finn	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21*, 109-141.
German	Isabel	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21,* 109-141.
German	Jores	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21,* 109-141.
German	Kerstin	Miller, M. (1979). The logic of language development in early childhood. Berlin: Springer-Verlag.
German	Konstantin	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21*, 109-141.
German	Leo	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21*, 109-141.
German	Leo	Behrens, Heike (2006). The input-output relationship in first language acquisition. *Language and Cognitive Processes, 21,* 2-24.
German	Lisa	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21*, 109-141.
German	Leon	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21*, 109-141.
German	Neele	Szagun, G. (2001). Learning different regularities: The acquisition of noun

		plurals by German-speaking children. *First Language, 21*, 109-141.
German	Pauline	heike.behrens@unibas.ch
German	Rahel	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21*, 109-141.
German	Sebastian	heike.behrens@unibas.ch
German	Sina	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language, 21*, 109-141.
German	Simone	Miller, M. (1979) The logic of language development in early childhood. Springer-Verlag, Berlin.
German	Soren	Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children*. First Language, 21*, 109-141.
Greek	Maria	Doukas, T. & Marinis, T. (2012). The acquisition of person and number morphology within the verbal domain in early Greek. *University of Reading Language Studies Working Papers, Vol. 4.*
Hebrew	Hagar	Arnon-Lotem, S. (1997). The Minimalist child: Parameters and functional heads in the acquisition of Hebrew. Unpublished doctoral dissertation, Tel Aviv University.
Hebrew	Leor	Arnon-Lotem, S. (1997). The Minimalist child: Parameters and functional heads in the acquisition of Hebrew. Unpublished doctoral dissertation, Tel Aviv University.
Hebrew	Lior	Arnon-Lotem, S. (1997). The Minimalist child: Parameters and functional heads in the acquisition of Hebrew. Unpublished doctoral dissertation, Tel Aviv University.
Hebrew	Ruti	No citation provided
Hebrew	Sivan	No citation provided
Hebrew	Smadar	Arnon-Lotem, S. (1997). The Minimalist child: Parameters and functional heads in the acquisition of Hebrew. Unpublished doctoral dissertation, Tel Aviv University.
Hungarian	Eva	MacWhinney, B. (1974). How Hungarian children learn to speak. Unpublished doctoral dissertation, University of California, Berkeley.
Hungarian	Gyuri	MacWhinney, B. (1974). How Hungarian children learn to speak. Unpublished doctoral dissertation, University of California, Berkeley.
Hungarian	Miki	No citation provided
Hungarian	Zoli	MacWhinney, B. (1974). How Hungarian children learn to speak. Unpublished doctoral dissertation, University of California, Berkeley.
Indonesian	Hizkia	Gil, D., & Tadmor, U. (2007). The MPI-EVA Jakarta child language database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.
Indonesian	Ido	Gil, D., & Tadmor, U. (2007). The MPI-EVA Jakarta child language database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.
Indonesian	Larissa	Gil, D., & Tadmor, U. (2007). The MPI-EVA Jakarta child language database. A joint project of the Department of Linguistics, Max Planck Institute for

		Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.
Indonesian	Michael	Gil, D., & Tadmor, U. (2007). The MPI-EVA Jakarta child language database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.
Indonesian	Pipit	Gil, D., & Tadmor, U. (2007). The MPI-EVA Jakarta child language database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.
Indonesian	Priska	Gil, D., & Tadmor, U. (2007). The MPI-EVA Jakarta child language database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.
Indonesian	Rizka	Gil, D., & Tadmor, U. (2007). The MPI-EVA Jakarta child language database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.
Indonesian	Timothy	Gil, D., & Tadmor, U. (2007). The MPI-EVA Jakarta child language database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.
Irish	Eoin	Cameron-Faulkner, T., & Hickey, T. (2011). Form and function in Irish child directed speech. Cognitive Linguistics, 22, 569-594. DOI:10.1515/COGL.2011.022.
Italian	Cam	Antelmi, Donna (1997). La prima grammatica del l'italiano: Indagine longitudinales ull'acquisizione del la morfosintassi italiana. Mulino, Bologna.
Italian	Diana	Cipriani, P., Pfanner, P., Chilosi, A., Cittadoni, L., Ciuti, A., Maccari, A., Pantano, N., Pfanner, L., Poli, P., Sarno, S., Bottari, P., Cappelli, G., Colombo, C., & Veneziano, E. (1989). *Protocollidiagnostici e terapeuticinellosviluppo e nellapatologia del linguaggio.* Italian Ministry of Health: Stella Maris Foundation.
Italian	Guglielmo	Cipriani, P., Pfanner, P., Chilosi, A., Cittadoni, L., Ciuti, A., Maccari, A., Pantano, N., Pfanner, L., Poli, P., Sarno, S., Bottari, P., Cappelli, G., Colombo, C., & Veneziano, E. (1989). *Protocollidiagnostici e terapeuticinellosviluppo e nellapatologia del linguaggio.* Italian Ministry of Health: Stella Maris Foundation.
Italian	Marco	Tonelli, L., Dressler, W. U., Romano, R. (1995). Frühstufen des Erwerbs der italienischen Konjugation. *Suvremena Linguistika 21,* 3-15.
Italian	Martina	Cipriani, P., Pfanner, P., Chilosi, A., Cittadoni, L., Ciuti, A., Maccari, A., Pantano, N., Pfanner, L., Poli, P., Sarno, S., Bottari, P., Cappelli, G., Colombo, C., & Veneziano, E. (1989). *Protocollidiagnostici e terapeuticinellosviluppo e nellapatologia del linguaggio.* Italian Ministry of Health: Stella Maris Foundation.
Italian	Raffaelo	Cipriani, P., Pfanner, P., Chilosi, A., Cittadoni, L., Ciuti, A., Maccari, A., Pantano, N., Pfanner, L., Poli, P., Sarno, S., Bottari, P., Cappelli, G., Colombo,

		C., & Veneziano, E. (1989). *Protocollidiagnostici e terapeuticinellosviluppo e nellapatologia del linguaggio.* Italian Ministry of Health: Stella Maris Foundation.
Italian	Rosa	Cipriani, P., Pfanner, P., Chilosi, A., Cittadoni, L., Ciuti, A., Maccari, A., Pantano, N., Pfanner, L., Poli, P., Sarno, S., Bottari, P., Cappelli, G., Colombo, C., & Veneziano, E. (1989). *Protocollidiagnostici e terapeuticinellosviluppo e nellapatologia del linguaggio.* Italian Ministry of Health: Stella Maris Foundation.
Italian	Viola	Cipriani, P., Pfanner, P., Chilosi, A., Cittadoni, L., Ciuti, A., Maccari, A., Pantano, N., Pfanner, L., Poli, P., Sarno, S., Bottari, P., Cappelli, G., Colombo, C., & Veneziano, E. (1989). *Protocollidiagnostici e terapeuticinellosviluppo e nellapatologia del linguaggio.* Italian Ministry of Health: Stella Maris Foundation.
Japanese	Aki	Miyata, S. (1992). Wh-questions of the third kind: The strange use of wa-questions in Japanese children. *Bulletin of Aichi Shukutoku Junior College No. 31* (pp. 151-155).
Japanese	Asato	Oshima-Takane, Y. & MacWhinney, B. (1995). CHILDES Manual for Japanese. Montreal: McGill University / Nagoya: Chukyo University.
Japanese	Arika	Oshima-Takane, Y. & MacWhinney, B. (1995). CHILDES Manual for Japanese. Montreal: McGill University / Nagoya: Chukyo University.
Japanese	Ishii	Ishii, T. (1999), The JUN Corpus, Unpublished.
Japanese	Nanami	Oshima-Takane, Y. & MacWhinney, B. (1995). CHILDES Manual for Japanese. Montreal: McGill University / Nagoya: Chukyo University.
Japanese	Noji	Noji, Junya. (1973-77). Yooji no gengoseikatsu no jittai I-IV. Bunka Hyoron Shuppan.
Japanese	Ryo	Miyata, S. (1992). Wh-questions of the third kind: The strange use of wa-questions in Japanese children. *Bulletin of Aichi Shukutoku Junior College No. 31* (pp. 151-155).
Japanese	Tai	Miyata, S. (1992). Wh-questions of the third kind: The strange use of wa-questions in Japanese children. *Bulletin of Aichi Shukutoku Junior College No. 31* (pp. 151-155).
Japanese	Taro	Hamasaki, N. (2002). The timing shift of two-year-olds' responses to caretakers' yes/no questions. In: Shirai, Y., Kobayashi, H., Miyata, S., Nakamura, K., Ogura, T. & Sirai, H. (Eds.), *Studies in language sciences: Papers from the Second Annual Conference of the Japanese Society for Language Sciences* (pp. 193-206).
Japanese	Tomito	Oshima-Takane, Y. & MacWhinney, B. (1995). CHILDES Manual for Japanese. Montreal: McGill University / Nagoya: Chukyo University.
Korean	Jiwon	No citation provided
Mandarin	XiXi	Tardif, T. (1993). Adult-to-child speech and language acquisition in Mandarin Chinese. Unpublished doctoral dissertaion, Yale University.
Mandarin	HaoYu	Tardif, T. (1993). Adult-to-child speech and language acquisition in Mandarin Chinese. Unpublished doctoral dissertaion, Yale University.
Mandarin	LiChen	Tardif, T. (1993). Adult-to-child speech and language acquisition in Mandarin Chinese. Unpublished doctoral dissertaion, Yale University.

Mandarin	LinLin	Tardif, T. (1993). Adult-to-child speech and language acquisition in Mandarin Chinese. Unpublished doctoral dissertation, Yale University.
Mandarin	BingBing	Tardif, T. (1993). Adult-to-child speech and language acquisition in Mandarin Chinese. Unpublished doctoral dissertation, Yale University.
Mandarin	"WX"	Tardif, T. (1993). Adult-to-child speech and language acquisition in Mandarin Chinese. Unpublished doctoral dissertation, Yale University.
Mandarin	YangYang	Tardif, T. (1993). Adult-to-child speech and language acquisition in Mandarin Chinese. Unpublished doctoral dissertation, Yale University.
Polish	Basia	Smoczynska, M. (1985). The acquisition of Polish. In D. Slobin (ed.), *The cross-linguistic study of language acquisition*, *Vol. 3* (pp. 595-686). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Polish	Inka	Smoczynska, M. (1985). The acquisition of Polish. In D. Slobin (ed.), *The cross-linguistic study of language acquisition*, *Vol. 3* (pp. 595-686). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Polish	Jadzia	Smoczynska, M. (1985). The acquisition of Polish. In D. Slobin (ed.), *The cross-linguistic study of language acquisition*, *Vol. 3* (pp. 595-686). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Polish	Janeczek	Smoczynska, M. (1985). The acquisition of Polish. In D. Slobin (ed.), *The cross-linguistic study of language acquisition*, *Vol. 3* (pp. 595-686). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Polish	Jas	Smoczynska, M. (1985). The acquisition of Polish. In D. Slobin (ed.), *The cross-linguistic study of language acquisition*, *Vol. 3* (pp. 595-686). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Polish	Kasia	Smoczynska, M. (1985). The acquisition of Polish. In D. Slobin (ed.), *The cross-linguistic study of language acquisition*, *Vol. 3* (pp. 595-686). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Polish	Krzys	Smoczynska, M. (1985). The acquisition of Polish. In D. Slobin (ed.), *The cross-linguistic study of language acquisition*, *Vol. 3* (pp. 595-686). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Polish	Marta	Weist, Richard, &Witkowska-Stadnik, Katarzyna. (1986). Basic relations in child language and the word order myth. International Journal of Psychology, 21, 363–381.
Polish	Michal	Smoczynska, M. (1985). The acquisition of Polish. In D. Slobin (ed.), *The cross-linguistic study of language acquisition*, *Vol. 3* (pp. 595-686). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Polish	Tenia	Smoczynska, M. (1985). The acquisition of Polish. In D. Slobin (ed.), *The cross-linguistic study of language acquisition*, *Vol. 3* (pp. 595-686). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Polish	Wawrzon	Weist, Richard, &Witkowska-Stadnik, Katarzyna. (1986). Basic relations in child language and the word order myth. *International Journal of Psychology, 21,* 363–381.
Portuguese	Gabriel	Guimarães, A. M. (1994). Desenvolvimento da linguagem da criançanafase del etramento. *Cadernos de Estudos Linguísticos, 26*, 103–110.
Portuguese	Paulo	No citation provided
Romanian	Bianca	Avram, Larisa (2001) Early omission of articles in child Romanian and the emergence of DP. *Revue Roumaine de Linguistique XLVI* (pp. 105-123).

Russian	Tanja	Bar-Shalom, E., & Snyder, W. (1997). Optional infinitives in Russian and their implications for the pro-drop debate. In M. Lindseth and S. Franks (eds.) *Formal approaches to Slavic linguistics: The Indiana Meeting 1996* (pp.38–47). Ann Arbor: Michigan Slavic Publications.
Russian	Varv	No citation provided
Sesotho	Hlobohang	Demuth, K. (1992). Acquisition of Sesotho. In D. Slobin (ed.), *The crosslinguistic study of language acquisition: Vol. 3* (pp. 557-638). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Sesotho	Litlhare	Demuth, K. (1992). Acquisition of Sesotho. In D. Slobin (ed.), *The crosslinguistic study of language acquisition: Vol. 3* (pp. 557-638). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Sesotho	Tsebo	Demuth, K. (1992). Acquisition of Sesotho. In D. Slobin (ed.), *The crosslinguistic study of language acquisition: Vol. 3* (pp. 557-638). Hillsdale, N.J.: Lawrence Erlbaum Associates.
Spanish	Idaira	No citation provided
Spanish	Irene	No citation provided
Spanish	Juan	Linaza, J., Sebastián, M. E., & del Barrio, C. (1981). Lenguaje, comunicación y comprensión. La adquisición del lenguaje. *Monografía de Infancia y Aprendizaje, 195-198.*
Spanish	Lucia	Aguado-Orea, J. & Pine, J. M. (2015). Comparing different models of the development of verb inflection in early child Spanish. *PLoS ONE 10: e0119613. doi:10.1371/journal.pone.0119613*
Spanish	Magin	Aguirre, C., (2000). La adquisición de las categorías gramaticales en español. Ediciones de la Universidad Autónoma de Madrid.
Spanish	Koki	Montes, R. (1987). Secuencias de clarificación en conversaciones con niños (Morphe 3-4): Universidad Autónoma de Puebla.
Spanish	Juan	Aguado-Orea, J. & Pine, J. M. (2015). Comparing different models of the development of verb inflection in early child Spanish. *PLoS ONE 10: e0119613. doi:10.1371/journal.pone.0119613*
Spanish	Maria	López-Ornat, S. 1994. *La adquisición de la lengua Española*. Madrid: Siglo XXI.
Spanish	Rafael	No citation provided
Spanish	Emilio	Vila, I. (1990) *Adquisición y desarrollo del lenguaje.* Barcelona: Graó.
Spanish	Yasmin	No citation provided
Swedish	Anton	Plunkett, K. & Strömqvist, S. (1992). The acquisition of Scandinavian languages. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol. 3* (pp. 457-556). Hillsdale, NJ: Lawrence Erlbaum Associates.
Swedish	Bella	Plunkett, K. & Strömqvist, S. (1992). The acquisition of Scandinavian languages. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol. 3* (pp. 457-556). Hillsdale, NJ: Lawrence Erlbaum Associates.
Swedish	Harry	Plunkett, K. & Strömqvist, S. (1992). The acquisition of Scandinavian languages. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol. 3* (pp. 457-556). Hillsdale, NJ: Lawrence Erlbaum Associates.
Swedish	Markus	Plunkett, K. & Strömqvist, S. (1992). The acquisition of Scandinavian languages. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol. 3* (pp. 457-556). Hillsdale, NJ: Lawrence Erlbaum Associates.

Swedish	Tea	Plunkett, K. & Strömqvist, S. (1992). The acquisition of Scandinavian languages. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol. 3* (pp. 457-556). Hillsdale, NJ: Lawrence Erlbaum Associates.
Tamil	Vanitha	No citation provided
Welsh	Alaw	Aldridge, M., Borsley, R. D., Clack, S., Creunant, G., & Jones, B. M. (1998). The acquisition of noun phrases in Welsh. In *Language acquisition: Knowledge representation and processing. Proceedings of GALA '97*. Edinburgh: University of Edinburgh Press.
Welsh	Bethan	Aldridge, M., Borsley, R. D., Clack, S., Creunant, G., & Jones, B. M. (1998). The acquisition of noun phrases in Welsh. In *Language acquisition: Knowledge representation and processing. Proceedings of GALA '97*. Edinburgh: University of Edinburgh Press.
Welsh	Dewi	Aldridge, M., Borsley, R. D., Clack, S., Creunant, G., & Jones, B. M. (1998). The acquisition of noun phrases in Welsh. In *Language acquisition: Knowledge representation and processing. Proceedings of GALA '97*. Edinburgh: University of Edinburgh Press.
Welsh	Melisa	Aldridge, M., Borsley, R. D., Clack, S., Creunant, G., & Jones, B. M. (1998). The acquisition of noun phrases in Welsh. In *Language acquisition: Knowledge representation and processing. Proceedings of GALA '97*. Edinburgh: University of Edinburgh Press.
Welsh	Rhian	Aldridge, M., Borsley, R. D., Clack, S., Creunant, G., & Jones, B. M. (1998). The acquisition of noun phrases in Welsh. In *Language acquisition: Knowledge representation and processing. Proceedings of GALA '97*. Edinburgh: University of Edinburgh Press.
Welsh	Rhys	Aldridge, M., Borsley, R. D., Clack, S., Creunant, G., & Jones, B. M. (1998). The acquisition of noun phrases in Welsh. In *Language acquisition: Knowledge representation and processing. Proceedings of GALA '97*. Edinburgh: University of Edinburgh Press.